

全国统计教材编审委员会推荐使用教材（2003 年第 2 版）

# SPSS 统计分析

## (第 5 版)

卢纹岱 朱红兵 主编  
吴喜之 审校

電子工業出版社  
Publishing House of Electronics Industry  
北京 · BEIJING

## 内 容 简 介

《SPSS 统计分析（第 5 版）》是在前 4 版的基础上，根据读者的反馈意见重新编写的。软件版本基于 SPSS 20 中文版。全书内容以统计分析应用为主，简要介绍各种统计分析方法的基本思想和基本概念；详细叙述操作方法，每种分析方法均给出对应的例题，例题涉及各个领域。每个例题均从方法选择、数据文件结构、操作步骤和结果分析方面给予说明。本书保留前 4 版的统计分析方法，对基本操作的内容、SPSS 过程语句介绍及生成统计图形方面的内容进行了压缩，合并了部分章节，增加了自动线性建模、有序回归、二阶段最小二乘法、一般对数线性回归、Logit 对数线性回归、模型选择对数线性回归分析、新版非参数假设检验的界面及其使用方法等内容。为方便读者和减少篇幅，书中所有例题数据均按章节编号，并保存在华信教育资源网 [www.hxedu.com.cn](http://www.hxedu.com.cn)，读者可自行下载。本书另配有电子教案，向采纳本书作为教材的教师免费提供。

本书可作为高等院校统计计算与软件课程的本科生和研究生教材，也适合于从事分析和决策的社会各领域、各相关专业读者学习参考。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

### 图书在版编目(CIP)数据

SPSS 统计分析 / 卢纹岱, 朱红兵主编. — 5 版. — 北京: 电子工业出版社, 2015.4

统计分析教材

ISBN 978-7-121-24924-2

I. ①S… II. ①卢… ②朱… III. ①统计分析—软件包—高等学校—教材 IV. ①C819

中国版本图书馆 CIP 数据核字 (2014) 第 274687 号

策划编辑: 秦淑灵

责任编辑: 秦淑灵      文字编辑: 苏颖杰

印      刷:

装      订:

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱      邮编: 100036

开      本: 787×1092 1/16      印张: 45.25      字数: 1158 千字

版      次: 2000 年 6 月第 1 版

2015 年 4 月第 5 版

印      次: 2015 年 4 月第 1 次印刷

印      数: 4000 册      定价: 75.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话: (010) 88254888。

质量投诉请发邮件至 [zlt@phei.com.cn](mailto:zlt@phei.com.cn)，盗版侵权举报请发邮件至 [dbqq@phei.com.cn](mailto:dbqq@phei.com.cn)。

服务热线: (010) 88258888。

# 《SPSS 统计分析(第 5 版)》编委会

主 编： 卢纹岱 朱红兵

审 校： 吴喜之

副主编： 何丽娟 朱一力 沙 捷

编 委： 殷小川 梁 蕾 卢纹凯 张泰昌 刘建通 石国书

费青松 朱启钊 谢利辉 宋楚强 王 湛 贺芬兰

宋 峥 苏 林 盖文红 卢大存 崔梦晗 张 晨

宋 巖 陈 冬 朱江华 刘 瑶 席凯强 郭 娟

唐天齐 张 铮 唐 莉 任 静 崔 健

# 前 言

SPSS 软件原名为 Statistical Package for the Social Science，社会科学用统计软件包。2000 年 SPSS 公司将其英文全称改为“Statistical Product and Service Solutions”，意为“统计产品与服务解决方案”，是一个组合式软件包。它集数据整理、分析过程、结果输出等功能于一身，是世界著名的统计分析软件之一。

在我们推出本版本时，SPSS 已将其更名为“IBM SPSS Statistics 20”。

SPSS 使用 Windows 的窗口方式展示各种管理数据和分析方法的功能，使用对话框展示各种功能选择项，清晰、直观、易学易用，涵盖面广。读者只要掌握一定的 Windows 操作技能和统计分析原理，就可以使用该软件为特定的科研工作服务。即使统计学水平有限，也可以使用系统默认项得到初步的分析结果，从而免去了编写程序的复杂工作。由于它具有强大的图形功能，使用该软件不但可以得到分析后的数字结果，还可以得到直观、清晰、漂亮的统计图，形象地显示对原始数据和分析结果的各种描述。

SPSS 已经在我国的社会科学和自然科学的各个领域得到广泛应用并发挥了巨大作用。我们所编写的《SPSS 统计分析》第 1、2、3、4 版得到了广大读者的厚爱，成为受读者欢迎的畅销书，这就是一个很好的证明。

在前 4 版的基础上，我们编写了《SPSS 统计分析(第 5 版)》。

根据 SPSS 软件的发展和广大读者的要求，我们对原作进行了仔细的检查、修正与改写，并按照增加内容但不增加篇幅的原则做了如下的改动。

- (1) 本书软件操作内容适用于 SPSS 20 以上版本，兼顾 SPSS 19 以下版本。
- (2) 对基本操作的内容、SPSS 过程语句介绍及生成统计图形方面的内容进行了进一步压缩。
- (3) 将第 4 版中原第 18 章“多响应变量的分析”中的内容合并到第 6 章“构建表格”的 6.5 节“多重响应变量分析”中。

(4) 对于软件汉化中的名称与专业名称或习惯用法不一致之处，本书保留传统用法，因此，可能会出现图题中的标题与图中标题不一致的情形。对于明显汉化有误之处，书中也做了说明。

(5) 随着应用统计学知识的普及，并根据读者要求，相对于上一个版本，本书新加的内容主要有：

- 在第 11 章“回归分析”中，增加了自动线性建模、有序回归、两阶最小二乘法、最优尺度回归、一般对数线性回归、Logit 对数线性回归、模型选择对数线性回归分析；
- 在第 12 章“非参数分析”中增加了 12.9 节“新版非参数假设检验的界面及其使用方法”；
- 在第 17 章“时间序列分析”中增加了各种分析方法的算法。

本书共三大部分。

- 第 1 章至第 3 章主要介绍 SPSS 的基本操作、基本概念和操作环境的设置，以及利用软件的各种帮助功能自学的方法。
- 第 4 章至第 18 章主要介绍随机变量和分布函数的应用、日期时间的运算；描述统计方法和分析表格的生成方法。还详细介绍了均值比较与检验、方差分析(参数检验)、非参



数检验、相关分析、回归分析、聚类分析、判别分析、因子分析、对应分析、结合分析、时间序列分析、生存分析。

- 第 19 章和第 20 章详尽地介绍各种统计图形的生成、编辑、修饰的方法。

为便于初学者和非统计学专业的读者学习，本书章节的编排有利于读者由浅入深地系统学习统计学知识和正确选择分析方法。每章均对统计分析方法的基本思想或基本概念做了深入浅出的介绍；对软件的操作进行尽量详细的说明；并对每种分析方法配以相应的例题。本书各章节的例题从数据解释、数据文件结构、方法选择、软件操作、输出结果解释和结论等几方面加以详细的说明。本书大部分例题均为作者科研或教学中的实例，读者容易接受。

本书所有例题数据按章节编号保存在华信教育资源网 [www.hxedu.com.cn](http://www.hxedu.com.cn)，数据文件名均以“data”开头，接着是 2 位数字的章号，横线后是 2 位数字，表明数据文件在本章中出现的序号。文件类型主要是 SPSS 数据文件(dataxx-xx.sav)，也有少量 Excel 文件(dataxx-xx.xls)和文本文件(dataxx-xx.txt)。读者可以按照书中的数据清单（附录 B）查找并参照。为方便读者学习，每个分析方法的介绍除有些基本操作被简化外，基本彼此独立，读者可根据自己的需要自行安排阅读。

本书由卢纹岱、朱红兵主编，并特别邀请中国人民大学统计学院吴喜之教授审校，在此深表谢意！

本书各章编写情况如下：卢纹岱、张泰昌、宋峥、任静、卢大存、唐天齐、张晨完成了第 1~5 章；第 6 章由朱红兵、卢纹岱完成；何丽娟、崔健、崔梦晗完成了第 7 章；宋楚强、郭娟、唐莉、张铮完成了第 8~10 章；第 11 章由朱红兵、沙捷、刘瑶、盖文红共同完成；第 12 章由朱红兵、朱启钊、苏林共同完成；卢纹岱、陈冬共同完成第 13 章；卢纹岱、朱红兵、朱启钊合写了第 14 章，何丽娟、殷小川合写了第 15 章；第 16 章由卢纹岱、朱江华完成；第 17 章由朱红兵、朱启钊、刘建通、席凯强共同完成；第 18~20 章由朱一力、王湛、贺芬兰编写；全书的统稿及排版工作由卢纹岱、朱红兵负责。在编写过程中，金水高、卢纹凯、张泰昌教授、席凯强、刘建通、梁蕾副教授提供了部分例题数据。解利辉、石国书、费青松、王雁、席凯强、刘建通等老师在资料收集、数据录入、核对、利用 SPSS 软件绘图等方面做了大量工作，在此一并表示诚挚的感谢。

本书适用于从事数据分析或统计应用的各领域、各专业的研究人员、中高层管理人员和决策者，也可以作为要求掌握统计分析方法和 SPSS 软件操作的高等院校的本科生、研究生的教材和自学参考书。

为方便教学，本书另配有电子教案，向采纳本书作为教材的教师免费提供，可登录电子工业出版社华信教育资源网([www.hxedu.com.cn](http://www.hxedu.com.cn))或电话联系(010-88254531)获取。

由于水平有限，加之时间仓促，有待改进的地方仍然很多，不妥之处在所难免，恳请广大读者对本书继续提出批评指正，我们愿与各位同行和爱好者进行交流学习。反馈意见请发电子邮件至：

luwendai@tsinghua.org.cn

zhuhongbing@cipe.net.cn

helijuan@cipe.net.cn

zhuyili2008@sina.com

shajie@cipe.net.cn

编 者

# 目 录

第 1 章 SPSS 概述.....	1	第 2 章 数据与数据文件.....	40
1.1 软件安装与运行.....	1	2.1 变量定义与数据编辑.....	40
1.1.1 SPSS 软件安装方法.....	1	2.1.1 数据编辑器.....	40
1.1.2 SPSS 的启动与退出.....	1	2.1.2 定义变量.....	41
1.1.3 SPSS 运行管理方式.....	3	2.1.3 定义日期变量.....	45
1.2 窗口及其功能概述.....	3	2.1.4 数据录入与编辑.....	46
1.2.1 数据编辑窗口.....	4	2.1.5 根据已有的变量建立新变量.....	49
1.2.2 输出窗口.....	4	2.1.6 打开、保存与查看数据文件.....	52
1.2.3 语句窗口.....	5	2.2 数据文件的转换.....	54
1.2.4 【窗口】菜单.....	7	2.2.1 ASCII 码数据文件的转换.....	54
1.2.5 对话框及其使用方法.....	8	2.2.2 数据库文件的转换.....	62
1.2.6 设置工具栏中的工具图标 按钮.....	10	2.2.3 观测的查重.....	63
1.3 系统参数设置.....	11	2.3 数据文件操作.....	66
1.3.1 参数设置基本操作.....	11	2.3.1 数据文件的拆分与合并.....	66
1.3.2 常规参数设置.....	12	2.3.2 观测的排序与排秩.....	72
1.3.3 输出观察窗口参数设置.....	13	2.3.3 对变量值重新编码.....	74
1.3.4 数据属性参数设置.....	14	2.3.4 数据文件的转置与重新构建.....	78
1.3.5 货币变量自定义格式设置.....	16	2.4 观测的加权与选择.....	88
1.3.6 标签输出设置.....	16	2.4.1 定义加权变量.....	88
1.3.7 统计图形参数设置.....	18	2.4.2 选择参与分析的观测.....	89
1.3.8 输出表格参数设置.....	21	习题 2.....	90
1.3.9 文件默认存取位置设置.....	22	第 3 章 输出信息的编辑.....	92
1.3.10 缺失值处理.....	23	3.1 输出窗口中的文本浏览与编辑.....	92
1.4 统计分析功能概述.....	24	3.1.1 利用导航器浏览输出信息.....	92
1.5 数据与变量.....	25	3.1.2 编辑导航器中的输出项.....	94
1.5.1 常量与变量.....	25	3.2 输出表格中信息的编辑.....	95
1.5.2 操作符与表达式.....	27	3.2.1 表格编辑工具与常用编辑 方法.....	95
1.5.3 观测.....	28	3.2.2 表格的转置与行、列、层的 处理.....	98
1.5.4 SPSS 函数.....	29	3.2.3 表格外观的设置与编辑.....	100
1.6 获得帮助.....	36	3.2.4 输出信息的复制与打印.....	105
1.6.1 SPSS 帮助系统.....	36	习题 3.....	105
1.6.2 右键帮助.....	38		
习题 1.....	39		

<b>第 4 章 随机变量与分布函数的应用</b> ··· 106	<b>第 7 章 基本统计分析</b> ····· 161
4.1 随机变量与分布函数····· 106	7.1 频数分布分析····· 161
4.1.1 随机变量及其概率分布 ····· 106	7.1.1 频数分布分析过程 ····· 161
4.1.2 随机变量的函数 ····· 109	7.1.2 频数分布分析实例 ····· 163
4.2 随机变量与分布函数的应用 ··· 116	7.2 描述统计 ····· 166
4.2.1 符合分布要求的随机数的 生成 ····· 116	7.2.1 描述统计中的基本概念 ····· 166
4.2.2 概率密度函数与累积概率密 度函数的应用 ····· 118	7.2.2 描述统计分析过程 ····· 167
习题 4····· 121	7.2.3 描述统计分析实例 ····· 167
<b>第 5 章 日期和时间函数及其运算</b> ····· 122	7.3 探索分析 ····· 168
5.1 日期时间函数 ····· 122	7.3.1 探索分析的意义和数据要求 ··· 168
5.1.1 SPSS 日期时间概述····· 122	7.3.2 探索分析过程 ····· 170
5.1.2 日期时间常量与变量····· 122	7.3.3 探索分析实例 ····· 172
5.1.3 日期时间函数····· 124	7.4 交叉表分析····· 175
5.2 日期时间函数的应用 ····· 127	7.4.1 交叉表及其独立性卡方检验 的思路 ····· 175
5.2.1 日期时间型变量的格式转换·· 127	7.4.2 交叉表分析过程····· 176
5.2.2 日期时间型变量的算术运算·· 130	7.4.3 交叉表分析实例····· 179
习题 5····· 133	7.5 比率分析 ····· 182
<b>第 6 章 构建表格</b> ····· 134	7.5.1 比率分析过程 ····· 182
6.1 自定义表格 ····· 134	7.5.2 比率分析实例 ····· 184
6.1.1 自定义表格的概念 ····· 134	7.6 P-P 图和 Q-Q 图 ····· 185
6.1.2 自定义表格的操作 ····· 135	7.6.1 P-P 图和 Q-Q 图分析过程··· 185
6.2 汇总、统计指标与统计检验 ··· 136	7.6.2 P-P 图和 Q-Q 图分析实例··· 186
6.2.1 统计指标与汇总项 ····· 136	习题 7····· 188
6.2.2 表格中的统计检验 ····· 142	<b>第 8 章 均值比较与检验</b> ····· 189
6.3 标题与其他选项····· 142	8.1 均值比较与均值比较的检验··· 189
6.3.1 定义表格标题····· 142	8.1.1 均值比较的概念 ····· 189
6.3.2 定义表格选项····· 143	8.1.2 均值比较与检验的过程····· 189
6.4 自定义表格实例····· 144	8.2 均值过程 ····· 191
6.5 多响应变量的概念与分类 ····· 146	8.2.1 均值过程中的统计量 ····· 191
6.5.1 多响应变量的概念与分类 ··· 146	8.2.2 均值过程操作 ····· 192
6.5.2 定义与建立多响应变量集 ··· 148	8.2.3 分析实例····· 194
6.5.3 多响应变量的频数分布分析·· 149	8.3 单样本 T 检验 ····· 196
6.5.4 多响应变量的交叉表分析 ··· 153	8.3.1 单样本 T 检验的概念····· 196
6.5.5 使用表功能分析多响应 变量集····· 156	8.3.2 单样本 T 检验的实例····· 197
习题 6····· 160	8.4 独立样本 T 检验 ····· 198
	8.4.1 独立样本 T 检验的概念 ····· 198
	8.4.2 独立样本 T 检验的过程 ····· 199
	8.4.3 独立样本 T 检验的实例 ····· 199

8.5	配对样本 T 检验 .....	202		方差分析过程 .....	254
8.5.1	配对样本 T 检验的概念 .....	202	9.5.4	重复测量方差分析实例 .....	256
8.5.2	配对样本 T 检验的过程 .....	202	9.5.5	关于趋势分析 .....	259
8.5.3	配对样本 T 检验的实例 .....	203	9.6	方差成分分析 .....	262
习题 8	.....	204	9.6.1	方差成分分析过程 .....	263
第 9 章	方差分析 .....	205	9.6.2	方差成分分析实例 .....	265
9.1	方差分析的概念与方差分析 过程 .....	205	习题 9	.....	268
9.1.1	方差分析的概念 .....	205	第 10 章	相关分析 .....	269
9.1.2	方差分析中的术语 .....	207	10.1	相关分析的概念与相关分析 过程 .....	269
9.1.3	方差分析过程 .....	208	10.1.1	简单相关分析的概念 .....	269
9.2	单因素方差分析 .....	210	10.1.2	相关分析过程 .....	270
9.2.1	简单的一维方差分析 .....	210	10.2	两个变量间的相关分析 .....	271
9.2.2	单因素方差分析过程 .....	211	10.2.1	两个变量间的相关分析 过程 .....	271
9.2.3	单因素方差分析实例 .....	215	10.2.2	两个变量间的相关分析 实例 .....	272
9.3	单因变量多因素方差分析 .....	220	10.3	偏相关分析 .....	276
9.3.1	单因变量多因素方差分析 概述 .....	220	10.3.1	偏相关分析的概念 .....	276
9.3.2	单因变量多因素方差分析 过程 .....	220	10.3.2	偏相关分析过程 .....	277
9.3.3	随机区组设计的方差分析 实例 .....	226	10.3.3	偏相关分析实例 .....	277
9.3.4	2×2 析因试验方差分析实例 ..	228	10.4	距离分析 .....	280
9.3.5	拉丁方区组设计的方差分析 实例 .....	230	10.4.1	距离分析的概念 .....	280
9.3.6	协方差分析实例 .....	233	10.4.2	距离分析过程 .....	281
9.3.7	多维交互效应方差分析实例 ..	235	10.4.3	距离分析实例 .....	283
9.4	多因变量线性模型的方差 分析 .....	237	习题 10	.....	285
9.4.1	多因变量方差分析概述 .....	237	第 11 章	回归分析 .....	286
9.4.2	多因变量方差分析过程和 数据要求 .....	238	11.1	线性回归 .....	286
9.4.3	多因变量线性模型方差 分析实例 .....	240	11.1.1	一元线性回归 .....	286
9.5	重复测量设计的方差分析 .....	250	11.1.2	多元线性回归 .....	288
9.5.1	重复测量方差分析概述 .....	250	11.1.3	异常值、影响点、共线性 诊断 .....	290
9.5.2	重复测量方差分析的数据 文件结构 .....	253	11.1.4	变非线性关系为线性关系 ..	291
9.5.3	组内因素的设置与重复测量		11.1.5	线性回归过程 .....	292
			11.1.6	线性回归分析实例 .....	296
			11.1.7	自动线性建模 .....	301
			11.2	曲线估计 .....	314
			11.2.1	曲线回归概述 .....	314
			11.2.2	曲线回归过程 .....	314

11.2.3	曲线回归分析实例	315	11.10.1	最优尺度回归的概念	362
11.3	二项 Logistic 回归	317	11.10.2	最优尺度回归过程	372
11.3.1	Logistic 回归模型	317	11.10.3	最优尺度回归分析实例	378
11.3.2	二项 Logistic 回归过程	320	11.11	对数线性模型	381
11.3.3	二项 Logistic 回归分析实例	323	11.11.1	对数线性模型的概念	381
11.4	多分变量 Logistic 回归	326	11.11.2	一般对数线性回归分析	383
11.4.1	多分变量 Logistic 回归的概念	326	11.11.3	Logit 对数线性回归分析	391
11.4.2	多分变量 Logistic 回归过程	328	11.11.4	模型选择对数线性回归分析	398
11.4.3	多分变量 Logistic 回归分析实例	331	习题 11		406
11.5	有序变量 Logistic 回归	335	第 12 章	非参数检验	407
11.5.1	有序变量 Logistic 回归的概念	335	12.1	卡方检验	408
11.5.2	有序变量 Logistic 回归过程	337	12.1.1	卡方检验的基本概念	408
11.5.3	有序变量的 Logistic 回归分析实例	339	12.1.2	卡方检验过程	408
11.6	概率单位回归	342	12.1.3	卡方检验分析实例	410
11.6.1	概率单位回归的概念	342	12.2	二项分布检验	412
11.6.2	概率单位回归过程	343	12.2.1	二项分布检验的概念与操作	412
11.6.3	概率单位回归分析实例	344	12.2.2	二项分布检验分析实例	413
11.7	非线性回归	347	12.3	游程检验	413
11.7.1	非线性模型	347	12.3.1	游程检验的基本概念	413
11.7.2	非线性回归过程	349	12.3.2	游程检验过程	414
11.7.3	非线性回归分析实例	351	12.3.3	游程检验分析实例	415
11.8	加权回归	353	12.4	一个样本的柯尔莫哥洛夫-斯米诺夫检验	415
11.8.1	加权回归的概念	353	12.4.1	一个样本的柯尔莫哥洛夫-斯米诺夫检验的基本概念	415
11.8.2	加权回归过程	355	12.4.2	柯尔莫哥洛夫-斯米诺夫检验过程	416
11.8.3	加权回归分析实例	355	12.4.3	柯尔莫哥洛夫-斯米诺夫检验分析实例	417
11.9	二阶段最小二乘法	358	12.5	两个独立样本检验	417
11.9.1	二阶段最小二乘法的概念	358	12.5.1	两个独立样本检验的用途与基本操作	417
11.9.2	二阶段最小二乘法过程	359	12.5.2	两个独立样本检验分析实例	421
11.9.3	二阶段最小二乘法分析实例	360	12.6	多个独立样本检验	422
11.10	最优尺度回归	362	12.6.1	多个独立样本检验的用途与操作	422

12.6.2	多个独立样本检验分析 实例 .....	424	13.5.2	判别分析过程 .....	482
12.7	两个相关样本检验 .....	425	13.5.3	判别分析实例 .....	486
12.7.1	两个相关样本检验的用途 与操作 .....	425	13.5.4	逐步判别分析与实例 .....	493
12.7.2	两个相关样本检验分析 实例 .....	427	习题 13 .....		498
12.8	多个相关样本检验 .....	428	第 14 章	因子分析与对应分析 .....	499
12.8.1	多个相关样本检验的用途 与操作 .....	428	14.1	主成分分析与因子分析 .....	499
12.8.2	多个相关样本检验分析 实例 .....	429	14.1.1	主成分分析与因子分析 概述 .....	499
12.9	新版非参数假设检验的界面 及其使用方法 .....	430	14.1.2	因子分析过程 .....	504
12.9.1	单样本检验 .....	430	14.1.3	因子分析实例 .....	509
12.9.2	独立样本检验 .....	437	14.1.4	利用因子得分进行聚类 .....	512
12.9.3	相关样本检验 .....	443	14.1.5	市场研究中的顾客偏好 分析 .....	516
习题 12 .....		449	14.2	对应分析 .....	519
第 13 章	聚类分析与判别分析 .....	450	14.2.1	对应分析概述 .....	519
13.1	聚类分析、判别分析及其分析 过程 .....	450	14.2.2	对应分析过程 .....	520
13.1.1	聚类分析 .....	450	14.2.3	对应分析实例 .....	523
13.1.2	判别分析 .....	450	习题 14 .....		526
13.2	两步聚类 .....	451	第 15 章	信度分析与多维尺度分析 .....	527
13.2.1	两步聚类概述 .....	451	15.1	信度分析 .....	527
13.2.2	两步聚类过程 .....	452	15.1.1	信度分析的概念 .....	527
13.2.3	两步聚类分析实例 .....	455	15.1.2	信度分析过程 .....	530
13.3	快速聚类 .....	458	15.1.3	信度分析实例 .....	531
13.3.1	快速聚类概述 .....	458	15.2	多维尺度分析 (ALSCAL) .....	533
13.3.2	快速聚类过程 .....	458	15.2.1	多维尺度分析的功能与 数据要求 .....	533
13.3.3	快速聚类分析实例 .....	460	15.2.2	多维尺度分析过程 .....	533
13.4	系统聚类 .....	463	15.2.3	多维尺度分析实例 .....	536
13.4.1	系统聚类概述 .....	463	习题 15 .....		538
13.4.2	系统聚类过程 .....	464	第 16 章	结合分析 .....	539
13.4.3	样品系统聚类分析实例 .....	469	16.1	结合分析概述 .....	539
13.4.4	变量聚类概述 .....	476	16.2	正交试验设计 .....	540
13.4.5	变量聚类分析实例 .....	476	16.2.1	试验设计中的问题 .....	540
13.5	判别分析 .....	480	16.2.2	正交试验设计的思路 .....	540
13.5.1	判别分析概述 .....	480	16.2.3	正交试验设计过程 .....	542
			16.2.4	正交试验设计实例 .....	544
			16.2.5	正交设计过程语句 .....	546
			16.3	试验设计结果的打印 .....	551

16.3.1	设计结果打印过程	551	17.6	季节分解法	613
16.3.2	打印调查用卡片实例	552	17.6.1	季节分解法模型	613
16.3.3	正交试验设计打印过程 语句	553	17.6.2	季节分解法分析过程	615
16.4	结合分析的语句与编程	555	17.6.3	季节分解法分析实例	616
16.4.1	结合分析过程语句	555	17.7	频谱分析	616
16.4.2	结合分析语句实例	559	17.7.1	频谱分析概述	616
16.5	结合分析实例	562	17.7.2	频谱分析过程	619
16.5.1	课题分析与正交设计	562	17.7.3	频谱分析实例	620
16.5.2	调查准备与调查	564	17.8	互相关	621
16.5.3	结合分析编程与结果分析	566	17.8.1	互相关概述	621
习题 16		570	17.8.2	互相关过程	622
17.8.3	互相关实例	623	习题 17		624
第 17 章	时间序列分析	571	第 18 章	生存分析	625
17.1	时间序列的建立和平稳化	572	18.1	生存分析概述	625
17.1.1	缺失值数据的替换	572	18.1.1	生存分析与生存数据	625
17.1.2	建立时间序列新变量	573	18.1.2	生存时间函数	626
17.2	序列图	576	18.1.3	Kaplan-Meier 法	626
17.2.1	序列图过程	576	18.1.4	Cox 回归模型	627
17.2.2	序列图应用实例	577	18.1.5	Cox 依时协变量回归模型	627
17.3	建立时间序列模型	579	18.2	寿命表分析	628
17.3.1	指数平滑与 ARIMA 模型 概述	579	18.2.1	寿命表分析概述	628
17.3.2	选择分析变量	592	18.2.2	寿命表分析过程	628
17.3.3	选择统计量	597	18.2.3	寿命表分析实例	630
17.3.4	图表	599	18.3	Kaplan-Meier 分析	632
17.3.5	输出项目的过滤	600	18.3.1	Kaplan-Meier 分析概述	632
17.3.6	保存新变量	600	18.3.2	Kaplan-Meier 分析过程	632
17.3.7	建模的其他选项	601	18.3.3	Kaplan-Meier 分析实例	635
17.3.8	时间序列分析实例	602	18.4	Cox 回归风险比例模型分析	636
17.4	应用时间序列模型	605	18.4.1	Cox 回归分析概述	636
17.4.1	应用时间序列模型过程	606	18.4.2	Cox 回归分析过程	637
17.4.2	应用时间序列模型分析 实例	606	18.4.3	Cox 回归分析实例	640
17.5	自相关	607	18.5	Cox 依时协变量回归模型 分析	642
17.5.1	自相关系数与偏自相关 系数的计算	607	18.5.1	Cox 依时协变量回归分析 过程	642
17.5.2	自相关图	609	18.5.2	Cox 依时协变量回归分析 实例	643
17.5.1	自相关分析过程	610	习题 18		646
17.5.3	自相关分析实例	611			

第 19 章	生成统计图形 .....	648	19.7	帕累托图 .....	666
19.1	概述 .....	648	19.7.1	选择帕累托图类型 .....	667
19.2	条形图和 3D 条形图 .....	648	19.7.2	简单帕累托图 .....	667
19.2.1	选择图形类型 .....	649	19.7.3	堆栈帕累托图 .....	668
19.2.2	简单条形图 .....	649	19.8	控制图 .....	670
19.2.3	复式条形图 .....	652	19.8.1	选择控制图类型 .....	670
19.2.4	堆积面堆图 .....	652	19.8.2	平均值、极差、标准差 控制图 .....	670
19.2.5	3D 条形图 .....	653	19.8.3	单值-移动极差控制图 .....	672
19.3	线图、面积图、高低图和 圆图 .....	654	19.8.4	不合格品率、不合格品数 控制图 .....	673
19.3.1	选择图形类型 .....	654	19.8.5	变量缺陷数、单位缺陷数 控制图 .....	675
19.3.2	堆积面积图 .....	655	习题 19 .....		675
19.3.3	多线线图 .....	655	第 20 章	编辑统计图形 .....	676
19.3.4	垂线图 .....	656	20.1	认识图形组成 .....	676
19.3.5	简单高-低-闭合图 .....	656	20.2	编辑平面统计图 .....	677
19.3.6	聚类高低收盘图 .....	657	20.2.1	图形编辑途径 .....	677
19.3.7	简单极差图 .....	658	20.2.2	改变图形构成 .....	679
19.3.8	差分面积图 .....	659	20.2.3	图形与文字修饰 .....	685
19.3.9	饼图 .....	659	20.2.4	坐标轴的编辑 .....	687
19.4	箱图和误差条图 .....	660	20.2.5	图条的修饰 .....	690
19.4.1	选择箱图和误差条图类型 .....	660	20.2.6	图线的编辑 .....	691
19.4.2	简单箱图 .....	660	20.2.7	饼图编辑 .....	693
19.4.3	复式箱图 .....	660	20.2.8	散点图的编辑 .....	694
19.4.4	简单误差条图 .....	661	20.2.9	文件管理 .....	698
19.4.5	复式误差条图 .....	662	习题 20 .....		699
19.5	散点图 .....	663	附录 A	标准化、距离和相似性的计算 .....	700
19.5.1	选择散点图图式 .....	663	附录 B	数据清单 .....	706
19.5.2	简单散点图 .....	663	参考文献 .....		711
19.5.3	重叠散点图 .....	664			
19.5.4	矩阵散点图 .....	664			
19.5.5	简单点图 .....	665			
19.6	直方图 .....	666			



# 第 1 章 SPSS 概述

## 1.1 软件安装与运行

### 1.1.1 SPSS 软件安装方法

- (1) 开机，启动 Windows，将 SPSS 系统安装光盘放入光盘驱动器。
- (2) 启动 Windows 资源管理器，双击光盘驱动器图标，找到安装应用程序的 setup 图标，见图 1-1 (a)。双击该图标启动 SPSS 20，见图 1-1 (b)，自动转入安装程序的屏幕显示，见图 1-1 (c)。
- (3) 单击【下一步】按钮，系统自动进行软件包的解压缩工作，安装开始以后就可以按照屏幕提示一步步地进行操作，每步操作均要认真阅读屏幕显示的信息和提示。
- 当再次出现如图 1-1 (c) 所示画面，且按钮区出现【完成】按钮时，则单击该按钮，完成安装。

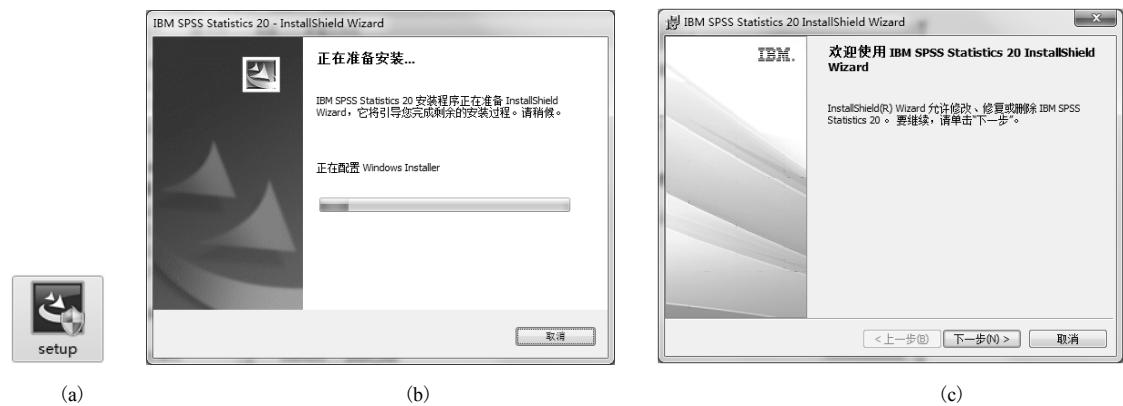


图 1-1 SPSS 20 的安装画面

### 1.1.2 SPSS 的启动与退出

#### 1. SPSS 的启动

- (1) 开机后，启动了 Windows，双击【开始】菜单的“IBM SPSS Statistics”选项，如图 1-2 (a) 所示。
- (2) 在提示画面后出现“IBM SPSS Statistics”对话框，如图 1-2 (b) 所示，共有 6 个功能选项和一个复选项。功能选项如下：
- 运行教程。选择此项打开如图 1-3 所示的教程窗口。可以根据主题单击书形图标，查看基本操作指导信息。单击右下角的箭头按钮翻页。
  - 输入数据。选择此项则显示数据编辑窗口，等待输入数据建立新数据集。
  - 运行现有查询。选择此项将显示打开文件窗口，从存储库检索需要的文件。读者可选择一个\*.spq 文件。

- 使用数据库向导创建新查询。选择此项打开如图 1-3 (b) 所示的数据库处理工具，将诸如 DBF 格式文件、XLS 格式的 Excel 文件、SQL 等数据库文件转换成 SPSS 数据文件。数据库处理工具的使用方法参见第 2 章。



图 1-2 选择菜单选项后出现对话框




图 1-3 运行教程

- 打开现有的数据源。选择此项读者可在第一个文件栏中选择一个.sav 格式的 SPSS 数据文件。
  - 打开其他文件类型。此项将让读者在第二个文件栏中选择一个其他格式的文件。例如选择一个常用的\*.spv，即 SPSS 的输出文件等。
- (3) 在对话框中单击【取消】按钮，跳过上述各项的选择，显示空数据编辑窗口【IBM SPSS Statistics 数据编辑窗口】，直接进入数据编辑状态，可以直接输入数据或操作菜单来打开已经存在的数据文件。
- (4) 如果在提示画面上选中【以后不再显示此对话框】，则下次启动 SPSS 时将不显示该对话框，而直接显示空数据编辑窗口。

2. SPSS 的退出

以下方法均可以退出 SPSS 系统。

- (1) 双击主画面左上角的窗口控制菜单图标,或单击该图标,在打开的小菜单中,单击【关闭】菜单项。
- (2) 单击主菜单的【文件】菜单项,在打开的【文件】菜单中,单击【退出】命令。
- (3) 单击数据编辑窗口右上角的图标。

1.1.3 SPSS 运行管理方式

1. 完全窗口菜单运行管理方式

SPSS 启动后即在屏幕上显示主画面,即数据编辑窗口,见图 1-4。完全窗口菜单管理方式指从数据输入、编辑、分析一直到分析结果的打印输出都在窗口中显示,通过菜单、对话框操作进行。

完全窗口运行管理方式主要在数据编辑窗口和输出窗口中进行操作。这种运行方式操作简便、直观,特别适用于初学者。由于窗口中包括的是基本参数和基本统计量的选项,因此完全窗口运行管理方式对某些专业人员来说,可能不能充分满足需要。

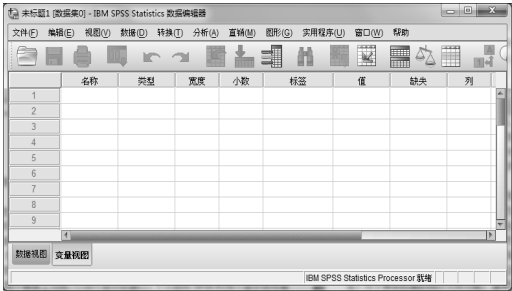


图 1-4 SPSS 数据编辑窗口

2. 程序运行管理方式

程序运行管理方式是在语句窗口中直接运行编写好的程序的一种方式。在该窗口中输入由 SPSS 命令组成的程序,利用键盘或主菜单中的【编辑】菜单项对窗口中的程序进行修改、编辑。在语句窗口中的程序可以分析数据窗口中的数据,也可以用有关的语句指定外部数据文件,对其进行分析,分析结果仍然显示在输出窗口中。习惯使用 SPSS 语言编写程序的读者仍然有用武之地。

3. 混合运行管理方式

混合运行管理方式是以上两种方法的结合。首先,在数据窗口中输入数据或利用【文件】菜单项打开已经存在的数据文件,然后利用对话框选择分析过程和分析参数。选择完成后不立即执行,而是用【粘贴】按钮将选择的过程及参数转换成相应的命令语句,置于语句窗口中。在该语句窗口中增加对话框中没有包括的语句和参数,或修改子命令中的参数,然后单击窗口中的【运行】功能按钮,将程序提交系统执行,结果显示在输出窗口中。混合运行管理方式既能简化操作,又可以弥补单纯窗口运行管理方式的不足。对于要求较高的统计分析功能,通常可使用这种方式。

1.2 窗口及其功能概述

SPSS 的文件系统包括 4 种基本类型的文件:Data(数据文件)、Syntax(语句文件)、Output(输出文件)和 Script(程序编辑文件)。每种类型的文件在各自的窗口中通过各自的菜单、功能按钮实现自己的各项功能。系统菜单的【文件】下拉菜单中的【新建】命令主要针对 4 个窗口中文件的操作,即当鼠标单击【文件】菜单中的【新建】命令打开小菜单时,显示可以新建各种类

型的文件：数据(文件)、语法(语句文件)、输出(文件)、脚本(文件)。对于使用 SPSS 的统计分析功能的读者来说，主要使用 3 种窗口，即数据窗口、输出窗口和语句窗口。

1.2.1 数据编辑窗口

SPSS 系统启动后激活该数据编辑窗口，如图 1-4 所示。未命名的数据编辑窗口最上方标有“未标题  $n$  [数据集  $m$ ] – IBM SPSS Statistics 数据编辑器”， $n$ 、 $m$  是打开窗口或数据文件的顺序号。窗口中有一个可扩展的平面二维表格，可以在此窗口中编辑数据文件。一旦保存了数据窗口中的数据，标题栏则显示该数据文件名。

对于数据窗口来说，无论“新建”还是“打开”命令，都会建立一个新的数据窗口。一次启动 SPSS 可以同时打开两个或两个以上的数据窗口。便于同时查看、操作两个以上的数据文件。单击标题栏激活数据编辑窗口。被激活的数据编辑窗口标题栏为蓝灰色(默认)，是当前工作窗；未被激活的数据编辑窗口标题栏是淡蓝色的。

1.2.2 输出窗口

SPSS 输出窗口标题栏中标有“输出 1 [文档 1] – IBM SPSS Statistics 输出查看器”，按照 SPSS 默认设置，输出窗口在启动后不显示在屏幕的主画面上。

1. 使用以下方法可以使输出窗口激活并显示在屏幕画面上

(1) 当使用了【分析】菜单中的统计分析功能处理数据窗口中的数据而产生输出信息时，输出窗口自动激活，显示在屏幕画面上。如果处理成功，则显示分析结果；如果处理过程中无法运行或发生错误，则在该窗口中显示系统给出的错误信息。

(2) 在【文件】菜单中选择“新建”项，在二级菜单中选择“输出”项，屏幕画面上显示一个输出窗口，见图 1-5。可以同时打开几个输出窗口，在窗口最上方的标题栏中按打开顺序显示窗口名：输出 1，输出 2，输出 3，…，在保存输出内容时由读者给出具体名称。

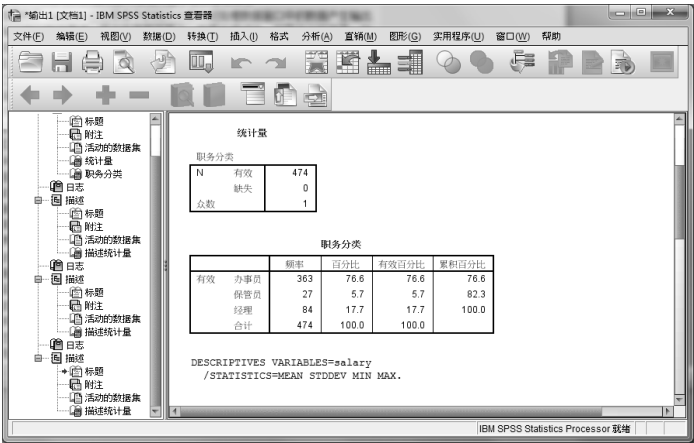


图 1-5 手动激活的输出窗口

2. 输出窗口组成




输出窗口除标题栏外，还包括以下几部分。

- 主菜单：由【文件】～【帮助】共 13 个菜单项组成。
- 工具栏：由各种功能的图标组成，是各种常用功能命令的快捷操作方式。

- 输出文本窗口：图标行下面右半边是一个文本窗口，在执行指定的操作或分析程序后，该窗口被激活，窗口显示输出信息，包括输出标题、文本、表格和统计图。该窗口中的内容可以利用鼠标、键盘和【编辑】菜单项的各种功能进行编辑。
- 输出导航窗口：导航窗口是浏览输出信息的导航器，位于图标行下面的左半边。以树形结构给出输出信息的提纲。
- 状态行：输出窗口的最下面一行是状态行，分为 4 个区，用鼠标指向任意一个区，就会在最左面区域显示每个区的功能解释。


### 3. 多个输出窗口的建立与主窗口的概念

单击【文件】菜单中的【新建】命令可以再打开一个输出窗口，打开的输出窗口按先后顺序标有输出 2，输出 3，…，过程执行结果只会输出到当前窗口（即工作输出窗口）。标记为工具栏中的灰色十字。其他输出窗口，即非当前输出窗口标记为工具栏中的绿色十字。

工作输出窗口（或称当前输出窗口）只能有一个。鼠标光标移到一个输出窗口中，单击该输出窗口中的图标按钮，就把该输出窗口激活为工作窗口，被激活的输出窗口的图标为灰色。所有操作，无论在哪个窗口单击【分析】菜单项进行的分析，都输出到被激活的窗口。非激活窗口的窗口工具栏中的图标是蓝色的.

单击【文件】菜单中的【新建】命令可以打开一个新的空输出窗口，打开命令把已经存在的输出文件显示到激活的输出窗口中。该输出窗口自动成为当前工作窗口。窗口标题栏显示文件名。

### 4. 关闭输出窗口

双击输出窗口左上角的图标，或单击输出窗口右上角的图标，都可关闭该输出窗口。如果窗口中的输出信息未存盘，则系统显示提示对话框。输出信息存盘后，窗口关闭。

### 5. 输出窗口能打开和保存的文件类型

输出窗口可以打开的文件类型有：Viewer document (\*.spv) 输出文件、Syntax (\*.sps) SPSS 语句文件、Draft Viewer document (\*.rft) 简化的输出文件、SPSS Script (\*.sbs) 脚本文件，以及无格式的 (\*.txt) 文本文件。文本文件和其他各类型文件只能在输出窗口中进行编辑。

## 1.2.3 语句窗口

### 1. 认识语句窗口

语句窗口由以下 5 部分组成，见图 1-6。

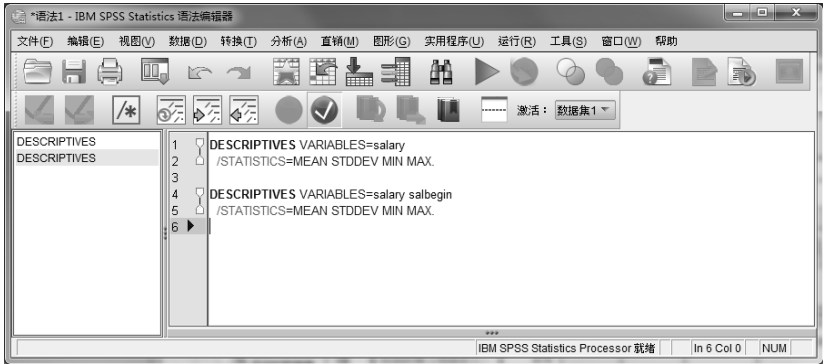



图 1-6 语句窗口

(1) 标题栏在窗口顶部, 标有“语法 n IBM SPSS statistics 语法编辑器”。

(2) 主菜单在标题栏下方, 包括【文件】~【帮助】共 12 个菜单项。

(3) 功能图标按钮在主菜单下方, 是可以简化操作的功能图标按钮, 包括打开文件、保存文件、打印文件、调用最近使用过的对话框、取消或重复执行用户上次操作、定位到数据(转向数据)、定位到观测(转向个案)、定位到变量、(转向变量)、显示变量信息(变量)、查找、运行选定内容、继续运行语句(继续运行语法)使用选择的变量集、显示所有变量、选择最后输出、显示语句帮助、是否主窗口标记等系统定制的图标按钮。

(4) 语句编辑区是图标下方的空白区域。在编辑区可以输入、编辑 SPSS 命令语句, 构成 SPSS 程序, 也可以输入和编辑文本文件。

(5) 状态行: 语句窗口也有状态行, 在窗口的最下面一行。

## 2. 语句窗口的激活与功能

(1) 打开一个语句窗口的方法与步骤:

① 单击主菜单的【文件】菜单项, 打开下拉菜单。

② 单击下拉菜单中的【新建】菜单项, 在二级菜单中单击【语法】项, 就打开了一个语句窗口, 如图 1-6 所示。

(2) 建立语句窗口的另一种方法是当选择了一种统计分析方法, 并在相应的对话框、子对话框中设置程序参数后, 在各可能生成命令程序的对话框中, 单击【粘贴】按钮, 自动打开一个语句窗口, 在语句窗口中生成与指定的统计分析及参数相应的 SPSS 命令语句。在语句窗口中可以对自动生成的命令语句进行编辑, 熟悉 SPSS 语句的读者可以增加对话框中不包括的参数或选项, 然后提交系统执行。

## 3. 语句窗口的结构

为便于程序人员编辑, 新的软件语句窗口分为 3 个子窗口:

最右面窗口显示整个程序, 包括全部命令语句、子命令和参数等。

最左面窗口只显示命令语句, 不显示子命令和参数; 顺序与右面窗口中命令出现的顺序一致。

中间窗口显示标记, 用沙漏图标显示鼠标光标所在的一个程序段, 从第一个语句到具有圆点的结束语句; 用红色圆球在窗口中间列显示程序分界。

## 4. 语句窗口的功能


(1) 各个 SPSS 过程的主对话框均有一个标有【粘贴】的图标按钮。它把 SPSS 过程的命令语句, 以及各选项对应的子命令语句, 按照 SPSS 语言的语法组成一个或若干个完整的程序并粘贴到主语句窗口中。


(2) 在语句窗口中可以使用键盘输入 SPSS 命令编写的 SPSS 程序, 每个过程语句(即一个完整的程序)均以圆点“.”结束。


用【编辑】菜单项中的各种功能和第二行图标按钮的功能编辑窗口中的程序。

用【文件】菜单项的各功能把窗口中的程序作为文件保存到磁盘中或关闭该窗口。

可以把已经存放在磁盘中的另一个程序文件调入, 独占该窗口或与已经存在于该窗口中的程序合并为一个程序作业, 以便合并一次运行。

(3) 当使用鼠标选择一个或几个完整的程序段后, 单击【运行】按钮, 就把该窗口中选中的程序提交系统执行。

(4) 单击【帮助】按钮，屏幕显示光标所在行上的命令或子命令所属的命令语句标准格式、可以选择的参数等，供读者查阅。

如果语句窗口中有多个过程语句，要执行其中的某一个过程，可以先用鼠标或键盘选择相应的语句，使之呈现反向显示，单击【运行】按钮即可提交系统执行。

4. 同时使用多个语句窗口

(1) 主语句窗口的概念

如果有几个实验程序，要放在几个语句窗口中时需要选择主语句窗口。

用前面介绍的从【文件】菜单选择【新建】的方法，可以同时打开若干语句窗口。在同一个 SPSS 期间，首先打开并粘贴了语句的语句窗口标为“\*语法 1”，第二个打开并粘贴了语句的语句窗口标为“语法 2”……但只能有一个主窗口。主语句窗口标题栏为蓝灰色，图标按钮为蓝灰色。主语句窗口的功能有别于非主语句窗口，各过程对话框中所选的选择项形成的命令语句和子命令组成的程序只能粘贴到主语句窗口中。非激活窗口的标题底色为天蓝色。主窗口图标按钮为蓝色。只有非主窗口的主窗口按钮(蓝色)可以操作，单击变为灰色，即变为主窗口。


(2) 选择主语句窗口

各对话框中的【粘贴】按钮产生的程序语句只生成在主语句窗口中。

① 单击“主窗口图标”，使其变为灰色。

② 当屏幕画面上有两个以上语句窗口时，使用鼠标单击窗口菜单，在下拉菜单中选择一个语句窗口，被选中的语句窗口变为主语句窗口。

(3) 关闭语句窗口

使用鼠标单击语句窗口左上角的 SPSS 语句窗口图标，选择下拉菜单中的【关闭】命令；使用复合控制键 Alt+F4；单击语句窗口右上角的图标，或选择【文件】菜单项的【退出】命令，在确定窗口中内容已经保存后，都可以关闭当前语句窗口。

5. 语句窗口打开与保存的文件类型是 SPSS 程序文件(\*.sps)。

1.2.4 【窗口】菜单

用鼠标单击【窗口】菜单，打开一个下拉菜单，见图 1-7。

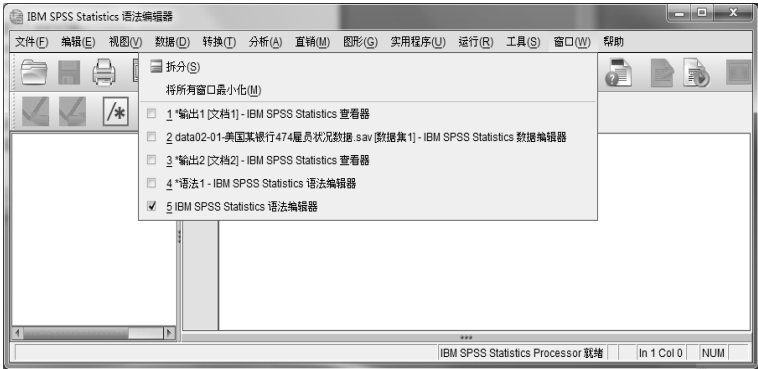


图 1-7 【窗口】菜单中的命令项

1. 选择窗口状态

单击【窗口】菜单中的【将所有窗口最小化】命令，当前所有窗口最小化，即变成几个图标按钮显示在 Windows 的状态栏内。

2. 各窗口之间的切换

在窗口菜单中列出了已经打开的窗口。打开的窗口名称前面，有对钩的窗口是主窗口，即当前工作窗口。没有对钩的窗口正处于非激活状态，如图 1-7 所示。

1.2.5 对话框及其使用方法

对话框，顾名思义就是提供人机对话环境和内容的窗口。主菜单中的各项命令基本上是通过对话框中的选项、复选项、变量、参数、语句等操作来实现的，通过对话框中的各种功能按钮展开下拉菜单、执行命令或打开子对话框。

1. 常见对话框类型

SPSS 中使用的对话框主要有如下 3 种。

(1) 文件操作对话框



例如打开已经存在的数据文件，按【文件】→【打开】→【数据】顺序逐一单击鼠标左键，打开【打开文件】对话框。与一般 Windows 应用软件的【打开文件】对话框不同的是，SPSS 的【打开文件】对话框有【粘贴】按钮，可以将打开文件的操作转换为命令语句粘贴到语句窗口中；而保存数据的窗口有选择要保存变量的功能。

(2) 统计分析主对话框

通过【分析】菜单中的各类统计分析命令所打开的第一个对话框，均为统计分析主对话框。在该对话框中选择参与分析的各类变量是该对话框的主要任务。另外，分析方法不同会有不同的其他选项，例如选择分析中的算法以及输出选项等。图 1-8 所示为相关分析的对话框。

SPSS 对话框中的变量表列出可以参与分析的变量标签，默认状态是变量名列在变量标签后面的中括号中。当变量标签与变量名太长，栏的宽度不够时，可以使用鼠标光标指向该变量所在的行，该变量的变量标签和变量名则显示在该行的加长区中。

如果在系统参数设置对话框【常规】选项卡的【变量】列表栏选择的是“显示名称”，则在【分析】对话框变量表中只显示变量名。可以使用鼠标右键单击任意一个变量名，在右键菜单中在【显示变量名称】与【显示变量标签】中选择一项，左侧变量表中只显示变量名，或者同时显示变量名与变量标签。

尺度变量使用“”黄色尺子在左边做标记。有序变量(也称等级变量或定序变量)的左端用“”三色彩球做标记，见图 1-8(a)。

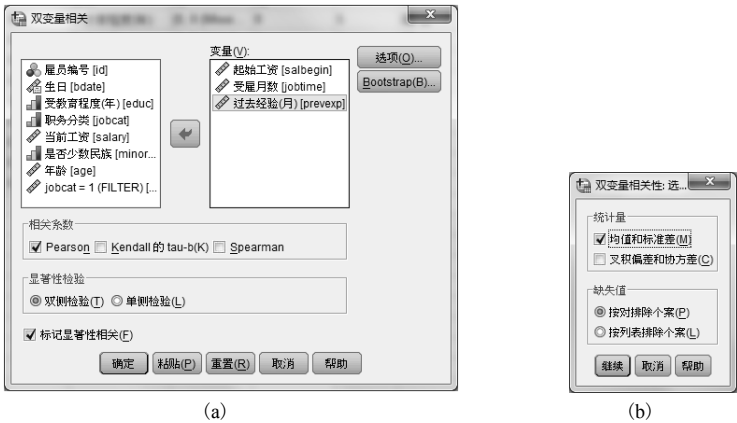


图 1-8 相关分析的主对话框和二级对话框



### (3) 其他选项对话框

其他选项对话框，即 SPSS 主菜单的其他菜单项对应的对话框或统计分析过程的二级对话框，这些对话框只在限定范围内提供选择的内容。图 1-8(b) 所示的对话框是相关分析的二级对话框。

## 2. 对话框的构成

### (1) 按钮

按钮的主要功能是激活选项。它告诉系统去做什么，包括以下 3 类，见图 1-9。

① 移动变量按钮，见图 1-9(b)。按钮中央是箭头，它把变量表中选中的变量加到变量框中。例如选择参与分析的变量，指定分类变量，或者指定因变量、自变量等。该按钮有时也用在构成模型时的变量选择。按钮的指向是可以改变的。当使用鼠标键选择了原始变量(左面矩形框中)时，箭头按钮指向右方，表示可以将选择的变量移到右边的变量表中去。当在右边的变量表中选择了变量时，箭头按钮指向左边，表示可以把变量表中的变量送回原始变量表中去，即从已经选择参与分析的变量中剔除。

② 打开下一级对话框的按钮，如图 1-9(c) 中的【选项】按钮，其特点是按钮中字符后面有省略号，按钮中的单词是下一级对话框的名称。这类按钮常用的还有：【模型】为打开建立模型对话框的按钮；【图形】(SPSS 中汉化为【绘制】)为打开作图对话框的按钮；【统计量】为打开统计量选择对话框的按钮；【保存】为保存新变量或保存新数据文件对话框的按钮等。



(a) (b) (c)

图 1-9 对话框中不同的功能按钮

③ 执行功能按钮，每个对话框中都有这样几个执行功能按钮。

- 【确定】按钮，见图 1-9(a)，单击这个按钮，把经过主菜单、子菜单、对话框，直到子对话框等选择的带有参数的命令过程语句提交系统执行。当选择或指定的变量、参数不符合运行相应过程的要求时，该按钮为灰色，不能提交系统运行。
- 【粘贴】按钮，鼠标单击该按钮，把通过对话框的各种操作组成的带有指定参数的过程命令语句显示到主语句窗口中。当选择或指定的变量、参数不符合运行相应过程的要求时，该按钮为灰色，表示没有具备生成可执行文件的条件。灰色按钮不能响应鼠标单击的操作。
- 【重置】按钮，清除在对话框中进行的一切选择和设置，使其恢复到系统默认状态。
- 【取消】按钮，取消本次打开对话框后的操作，返回到上一级对话框或窗口。
- 【帮助】按钮，打开帮助窗口，显示与当前对话框及其各项有关的帮助信息。
- 【继续】按钮，一般是二级对话框中的按钮。单击该按钮表明确认在二级对话框中的参数选择，返回前一级对话框。与之并列的有【取消】按钮和【帮助】按钮。

### (2) 选项

选项有两种。单选项形状像一个收音机旋钮，如图 1-10 所示。总是多个带有旋钮的选项排列在一起。这些选项只能择其一，不能同时选两个或两个以上。被选中的一项前面的圆圈旋钮中出现黑点。并列的若干项中必须且只能选择其中的一项。如果只有一项，无与之并列的项，则选择与否均可。

复选项形状为方框，被选中的复选项前有“√”出现，如图 1-11 所示。可以同时选中多个复选项，也可以一个不选。任何一项都不选时，有时会产生不希望产生的结果，或者输出窗口中没有分析结果输出，甚至出错。

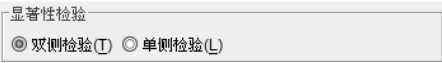


图 1-10 单选项

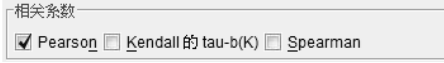


图 1-11 复选项

1.2.6 设置工具栏中的工具图标按钮

各窗口中都有自己的工具栏，工具栏中显示常用功能的图标按钮，这些图标按钮使许多操作变得简单方便。

如果当前窗口中没有这些工具图标按钮，可以使用下述方法将这些工具图标按钮显示在各窗口工具栏中。下面以编辑数据窗口中的工具栏为例，将“复制”、“剪切”、“删除”3 个编辑工具添加到工具栏中。操作步骤与方法如下：

(1) 在数据编辑窗口中，按【视图→工具栏→设定】顺序逐一单击鼠标左键，打开【显示工具栏】对话框，如图 1-12 所示。

(2) 在【窗口】框内，单击向下的箭头展开窗口表，由于每个窗口有不同的工具栏，要确定编辑哪个窗口的工具栏，就在窗口下拉列表中选择哪个窗口名。从数据窗的【视图】菜单启动【显示工具栏】对话框，【窗口框】中首先显示的是【数据编辑器】。

(3) 在【工具栏】内显示的是在【窗口】框中确定的窗口可以显示的工具栏名称。有的窗口可能同时出现两个以上工具栏名称选项，可以同时选择。但同时选择多个工具栏，会有重复的图标按钮出现在同一个窗口中。因此，最好使用系统默认的工具栏。

(4) 在工具栏内选择一个工具栏，使之显示彩色底纹。图 1-12 选择的是数据编辑器。如果需要建立全新的工具栏，则单击对话框右侧的【新建】按钮。

(5) 单击右边的【编辑...】按钮，打开【编辑工具栏】对话框，见图 1-13。对话框分为 3 个部分，左面【类别】栏列出的是当前窗口的菜单项。每一个菜单项对应一组工具图标。当选择了一个菜单项时，所对应的所有工具图标显示在右面的【工具】栏内。下面的【设定工具栏：数据编辑器】指定窗口的工具栏，它包括若干工具图标，是可以编辑的。







图 1-12 【显示工具栏】对话框



图 1-13 【编辑工具栏】对话框

(6) 在【类别】栏内选择一类工具，选择【编辑】工具，在右边的工具栏内显示编辑类的所有工具图标。

(7) 在对话框下边的【设定工具栏：数据编辑器】中，工具栏为当前该工具栏的现状，见图 1-14(a)。可以用鼠标将某个图标拖曳到新的位置，重新安排图标的排列。

(8) 在【工具】栏内，选择一个图标按钮。按下鼠标左键，拖曳到下边的工具栏中，松开鼠标键，选中的图标按钮出现在工具栏中。用这样的方法将图标 剪切(T)、 删除(D)、 复制(C) 一一拖曳到下面的工具栏中 图标按钮前边，见图 1-14 (b)。

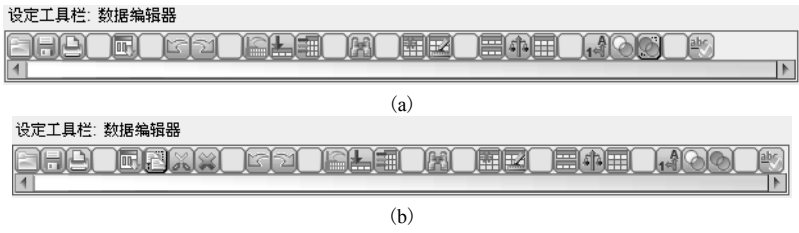


图 1-14 编辑前后的【数据编辑器】工具栏

(9) 单击【继续】按钮返回到如图 1-12 所示的【显示工具栏】对话框中。单击【确定】按钮，结束操作。此时数据窗口中的工具栏已经增加了 3 个工具图标。

(10) 在【编辑工具栏】对话框中编辑好的工具栏可以在其他窗口使用。不熟悉窗口、工具栏等操作的读者对此功能应慎重使用。如果需要重新安排，则单击该对话框下面的【重置工具栏】按钮，恢复定义之前的工具栏状态。

### 1.3 系统参数设置

#### 1.3.1 参数设置基本操作

系统初始状态和系统默认值的设置是通过【编辑】→【选项】对话框完成的，通过【编辑】菜单中的【选项】命令打开该对话框。参数与状态设置生效的时间不同，有的在确认后立即生效，有的则要在下次启动 SPSS 系统时才生效。但无论何时生效，只要生效，设定的状态或参数即代替了原来系统给定的默认值。

按【编辑】→【选项】顺序打开【选项】对话框，在如图 1-15 所示的【选项】对话框中进行系统状态和参数的设置。有以下几种可能的情况需要使用对话框执行功能按钮。



图 1-15 【选项】对话框【常规】选项卡

① 当完成任何参数或状态设置后,可单击**【确定】**按钮,确认所作的设置并返回到 SPSS 主画面。

② 如果在**【选项】**对话框中一系列设置完成后认为设置得不够满意,需要重新设置,可以单击**【取消】**按钮恢复到打开该对话框时的原始设置状态,重新进行设置工作。

③ 单击**【取消】**按钮退出**【选项】**对话框,返回到 SPSS 主画面,同时刚设置的参数作废。

④ 单击**【帮助】**按钮,打开与该对话框各项有关的帮助窗口,查看有关说明。

以上操作在每项设置过程中或完成后都可以进行,后续操作中类似设置不再重复。

### 1.3.2 常规参数设置

**【常规】**选项卡上可设置各种通用参数,见图 1-15。

#### 1. 设置显示变量、顺序的方式

**【常规】**选项卡上的左边第一栏是**【变量列表】**栏,下面的单选项设定变量在变量表中的显示方式和显示顺序,可通过两组选项进行选择。

##### (1) 变量的显示方式

① **【显示标签】**。选择此项,变量标签显示在前,变量名显示在后面的括号中。此为系统默认方式。

② **【显示变量名】**(SPSS 汉化为**【显示名称】**)。选择此项,在各对话框的源变量表中只显示变量名。

##### (2) 变量的显示顺序

① **【字母顺序】**,按变量名的字母顺序排列。

② **【文件】**,按变量在数据文件中出现的顺序排列。此为系统默认方式。

改变变量显示顺序的设置对当前的工作数据文件无效,只对选择应用和**【确定】**按钮以后打开或定义的数据文件起作用。在各统计分析对话框中,源变量表中的变量按选定的方式排列。

③ **【测量】**,按变量的测度水平名义、有序(SPSS 汉化为“序号”)、尺度(SPSS 汉化为“度量”)排列。

#### 2. 角色

在某些对话框中,如非参数假设检验的单样本、独立样本及相关样本等的对话框中,SPSS 过程可基于对变量定义的角色来预先选择变量的功能。

① **【使用预定义角色】**,默认情况下,基于对变量定义的角色,由程序自动选择变量。

② **【使用定制分配】**,由使用者自行设定变量的用途。

#### 3. 窗口状态的选择

**【常规】**选项卡左边第三栏标有 Windows,在其中选择窗口状态。

(1) **【观感】**下拉菜单中有 3 项,可以选择其中之一:

① **【SPSS Standard】**,使用 SPSS 标准窗口。

② **【SPSS Classic】**,使用 SPSS 经典界面。

③ Windows,一种具有表格线,颜色也不同的窗口。

(2) 在启动时打开语句窗口。习惯于使用 SPSS 语言编程和使用 SPSS 对话框功能的读者应该选择此项。

(3) 每次只打开一个数据集。选择此项，不能同时打开两个以上数据文件或数据窗口。

#### 4. 输出的设置

(1) 【表格中较小的数值没有科学计数法】。非常小的小数值将显示为 0 (或 0.000)。

(2) 【将本地数字分组格式应用到数值】。不同地区数值分组格式不同。我国使用逗号进行三位分隔数值。分组格式不适用于树、模型查看器、DOT 或 COMMA 格式的数值，以及 DOLLAR 或自定义货币格式的数值。但它适用于以 DTIME 格式数值来显示日期值，例如，以 ddd hh:mm 的格式显示 ddd 的值。

(3) 【测量系统】，在其下拉菜单中可以选择测度单位，即【磅】、【英寸】或【厘米】。它们的换算关系为：1 英寸 = 72 磅 = 2.54 厘米。如果需要作精细的图形，可以使用【磅】作为单位，系统默认单位为英寸。

(4) 【语言】选项，选择输出结果的默认语言，常用的除英文外还有：

① 【繁体中文】，输出使用繁体中文。如果没有安装繁体中文字库，不要轻易设置，否则结果会出现乱码。

② 【简体中文】。输出表格标题或输出项有时所用术语翻译有误。

无论选择哪一个，输出仍以英文为主，只在输出表格标题和输出项使用指定的语言。

(5) 【提示】栏控制在运行一个 SPSS 过程后在观察窗口中显示的输出结果的通告方式。有两个选项，默认同时使用两种方式。

① 【弹出浏览器窗口】。当有新处理结果时输出窗口自动弹出。

② 【滚动到新的输出】。当有新处理结果时屏幕显示到新的输出信息处。

(6) 【数据和语法的字符编码】栏，可用确定读写数据文件和语法文件默认的编码方式。

① 【Locale 的写入系统】。使用当前区域设置确定读写文件的编码方式。

② Unicode (通用字符编码)。使用通用字符集编码 (UTF-8) 来读写文件。

(7) 【用户界面】【语言】选择。默认【简体中文】。

### 1.3.3 输出观察窗口参数设置

在【查看器】选项卡上设置观察窗，即设置输出窗口的各种参数，见图 1-16。在改变参数设置后，单击【确定】按钮退出【选项】窗口后，再次运行 SPSS 命令，产生新的输出时才能生效。共有 4 部分参数可根据需要重新设置。

#### 1. 初始输出状态设置

在【查看器】选项卡的左边第一项，标有【初始输出状态】栏，设置各种输出的初始状态。

(1) 【项目图标】框与【项】下拉菜单中的项一一对应，可以在【项目图标】列选择，也可以在下拉菜单中选择，在该参数框中控制输出项在每次运行一个统计分析结果产生输出时，是否自动显示或隐藏，以及初始状态使用的对齐方式。可以选择的输出项有：【日志】、【警告信息】、【注释信息】(SPSS 汉化为【附注】)、【标题】、【页面标题】、【枢轴表】、【图表】、【文本输出信息】(表格中没有显示的输出信息)、【树形结构图】(SPSS 汉化为【树模型】)、【模型浏览器】。每选择一项，就可以按下面(2)、(3)项设定该项的状态。



图 1-16 【查看器】选项卡

1.3.4 数据属性参数设置

【数据】选项卡用来设置有关数据的各种参数，如图 1-17 所示。

(1) 【转换与合并选项】

SPSS 进行某些数据转换(如计算变量和重新编码为不同变量)时和文件转换(如增加变量或观测)后，可以不要求立即执行，而是到 SPSS 读取这些数据去执行另一个命令时再对数据或文件进行转换。何时执行转换，可以通过【转换与合并选项】下面的选项确定。

① 【立即计算值】，要求指定转换方法后立即执行。

② 【使用前计算值】，指定在使用之前再进行转换或合并。对一个大的数据文件，选择此项可以延迟执行以便节省处理时间。

(2) 【显示新数值变量的格式】

为新数值变量指定系统默认的显示宽度和小数位数。

如果一个数值相对于显示格式太长，则 SPSS 首先截掉小数部分，然后转化成科学计数法显示。显示格式对参与计算的数值本身没有影响，例如 123456.78 可以显示成 123456，但在进行任何计算时都使用未被截掉小数部分的原始值。

(2) 【初始内容】。在【初始内容】标题下面的单选项：【显示】、【隐藏】，确定项目列表中所指定的项目显示还是隐藏。

(3) 【调整】。指定文本内容的对齐方式。所有输出均默认左对齐，仅打印输出的对齐方式由【调整】项下面的单选项确定：【左对齐】、【居中对齐】(SPSS 汉化为【中间】)、【右对齐】。

(4) 选中最下面的【在日志中显示命令】选项，读者可以从日志中复制命令语句并将它们保存在一个语句文件中。

2. 标题、输出文本的字体、字号设置

在【查看器】选项卡右面有 3 个栏目：【标题】栏、【页面标题】栏、【文本输出】栏，分别定义各项的字体、字形、字号和颜色，这些设置对新产生的输出生效。

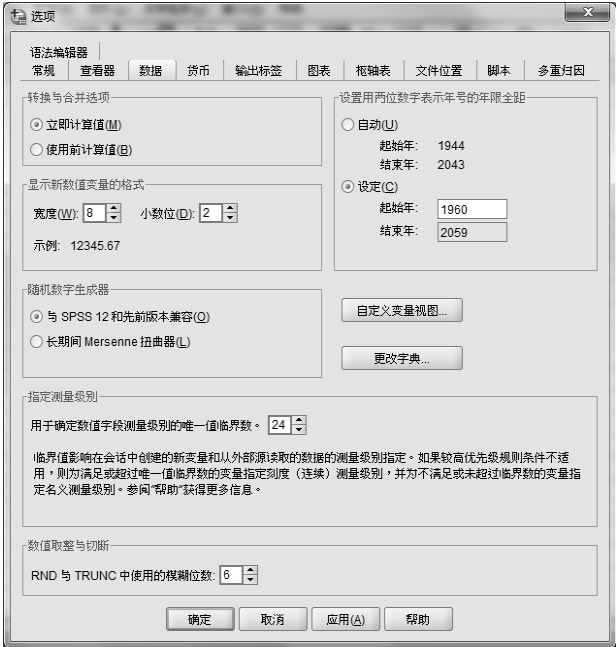


图 1-17 【数据】选项卡

- ① **【宽度】**。可以输入显示数值的总宽度。
- ② **【小数位】**。可以输入显示数值的小数位数。
- (3) **【随机数字生成器】**

有两个随机数生成器可以选择：

① **【与 SPSS 12 和先前版本兼容】**的随机数发生器。如果需要使用 12 版以前版本的随机数发生器产生使用指定种子数的随机数，就选择此项。

② **【长期间 Mersenne 扭曲器】**。是一种更可靠的新随机数发生器。

(4) **【设置用两位数字表示年号的年限全距】**

即对日期型数据中的年份指定使用两位数字输入和显示(例如 10/28/97、29-OCT-96)时：

① **【自动】**。自动指定表示年限范围项，根据当前年向前 69 年作为开始，向后 30 年作为结束。当前年即系统时间确定的年，加上当前的一年共 100 年的范围。如当前年是 2013 年，则自动设定的年限**【起始年】**为 1944 年，**【结束年】**是 2043 年。

② **【设定】**。自定义年限范围。读者可以输入年限范围的起始年。结束年参数框中的数值是系统自动确定并显示的。图 1-17 中，**【起始年】**输入“1960”，**【结束年】**显示为“2059”，范围也是 100 年。

(5) **【自定义变量视图...】**

单击该按钮，打开**【自定义变量视图】**对话框，如图 1-18 所示。在该对话框中重新安排和选择变量视图中的变量属性项。

- ① 选择一项，单击向上或向下箭头按钮，可以改变所选项的位置。
- ② 单击某项前的方框，若有对钩，则该项显示在数据视图中，没有对钩则不显示。
- (6) **【更改字典...】**

单击该按钮打开**【更改字典】**对话框，见图 1-19。在下拉列表中选择一种语言，该语言字典将用于在数据窗口中检查拼写，但是无中文字典可以选择。

(7) **【指定测量级别】**

在该栏中定义尺度数据与分类数据的界限。如果默认一个数值型变量至少有 24 个不同的数值，则认为它是尺度变量。不同数值的个数少于指定值，则系统认为是名义变量，或有序变量。可以单击上下箭头按钮增加或减少这个数值，更改这个参数。

此外，SPSS 系统默认货币或美元变量、日期时间(不包括月份、星期)变量，变量数值中至少有一个负数或至少包含一个非整数值为尺度变量，即连续变量。

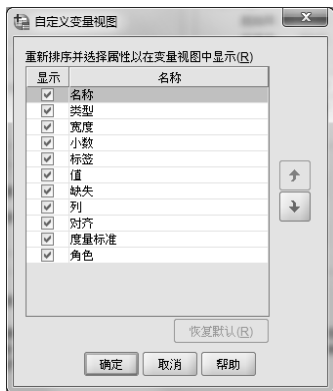


图 1-18 **【自定义变量视图】**对话框

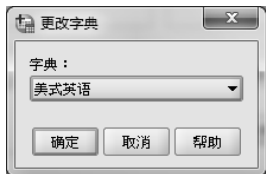


图 1-19 **【更改字典】**对话框

(8) 【数值取整与切断】

RND 和 TRUNC 中使用的模糊位数。此项设置这两个函数四舍五入取整或截尾取整的默认阈值。

1.3.5 货币变量自定义格式设置

SPSS 允许读者自己设定常用的货币数值型变量的输出格式，即显示格式。【货币】选项卡设置有关数据的各种参数，如图 1-20 所示。



图 1-20 【货币】选项卡

为“\$”。

(2) 【后缀】。设置在数值的后面添加的字符，系统默认值是空格。图 1-20 中设置为“/”。

3. 【负值】栏，设置负数的首尾字符

- (1) 【前缀】。在该框内，输入负数首字符，系统默认值是“-”。
- (2) 【后缀】。在该框内，输入尾字符，默认值是空格。图 1-20 中设置为“#”。

4. 【小数分隔符】栏，设置十进制数的小数点符号，同时确定三位的分隔符

- (1) 【句点】。用实心圆点作为小数点符号，每三位的分隔符为逗号。此为系统默认值。
- (2) 【逗号】。用半角逗号作为小数点符号，每三位的分隔符就自动设置为圆点。

以上参数设置完毕，按格式表达的数字样例显示在【样本输出】栏内。在图 1-20 中，上面一个是正数的样例，为\$1,234.56/；下面一个是负数的样例，为-\$1,234.56/#。选择格式的命名后即可单击【应用】按钮确认定义的格式。定义的格式即可在定义数值型变量时使用。

1.3.6 标签输出设置

输出结果或输出表格中，若将变量标签或变量值一并输出，能够让读者很方便地阅读这些结果和表格。这些变量标签或变量值标签都是在定义一个变量时，使用主菜单【数据】功能中的定义变量功能项定义的。【输出标签】选项卡（见图 1-21）是设定输出格式的。输出的表格显

最多可以创建 5 种自定义货币显示格式，其名称为 CCA、CCB、CCC、CCD 和 CCE。格式名称不能更改。格式可以包括特殊的前缀和后缀字符以及对负值的表示方式。

为自定义货币格式定义的前缀、后缀和小数指示符仅用于显示目的。不能用自定义货币字符在数据编辑器中输入值。

1. 【设定输入格式】栏，列出可以设置的自定义格式

5 种自定义格式的名称为 CCA、CCB、CCC、CCD 和 CCE。选择一个名称，例如图中选择 CCB，再做下面的操作。

2. 【所有值】栏，设置数值的首尾字符

(1) 【前缀】。设置在数值的前面添加的字符，系统默认值是空格。图 1-20 中设置



示变量名还是显示变量标签，遇到需要显示分类变量值时显示变量值还是显示它的值标签，可以使用【输出标签】选项卡中的两个栏目来设定。



图 1-21 【输出标签】选项卡

1. 在【轮廓标签】栏中，设定输出表格时在相应的导航栏中是否使用变量标签
- (1) 【项标签中的变量显示为】。该项设置变量标识。在下拉列表中有 3 个选项，指定其中一个。
- ① 【标签】。选择此项输出使用变量标签表示每个变量。

② 【名称】。选择此项输出使用变量名表示每个变量。

③ 【名称和标签】。选择此项同时使用变量名和变量标签表示每个变量。
- (2) 【项标签中的变量显示为】。在该栏中设置输出中变量值的表示方法。单击向下箭头按钮，在下拉列表中有 3 个选项，指定其中一个。
- ① 【标签】。选择此项使用变量值标签表示每个变量值。适用于分类变量的值。

② 【值】。选择此项直接输出变量值本身。

③ 【值和标签】。选择此项同时使用变量值和变量值标签表示每个变量的值。分类变量的输出可以选择此项。
2. 在【枢轴表标签】中设置表格标签
- 设定输出表格时是否使用标签，其操作过程与轮廓标签栏相同。
- 注意：当变量标签或值标签过长时，在图形或表格中使用标签不一定是合适的。因此，使用标签与否，要视实际情况而定。
- 输出标签选项只有对指定这些选项之后产生的输出生效，对当前已经在输出窗口中的输出图形或表格不起作用。

1.3.7 统计图形参数设置

【图表】选项卡用来设置统计图形的各种参数，如图 1-22 所示。



图 1-22 【图表】选项卡

1. 【图表模板】

设置图形模板，新图形可以套用这些参数。

- (1) 【使用当前设置】。使用当前系统默认的模板和此选项卡中的默认参数。
- (2) 【使用图表模板文件】。指定【使用图表模板文件】中设定的图形参数。选择此项，需要单击【浏览】按钮，在【打开】对话框中指定一个模板文件。

也可以建立新的模板文件，用需要的参数生成图形并将其保存到模板文件中。方法是生成图形后，双击图形，打开【图形编辑(Chart Editor)】对话框。在【文件】菜单中选择【保存】图形模板命令，设定保存的模板项目后把图形保存为扩展名为“.sgt”的模板文件。

2. 【图表宽高比】

设置图形的宽高比。

默认值为 1.25。在该参数框中可以直接输入需要的比例数值。输入的数值在 0.1~10.0 之间，设置比例小于 1，图形高度大于图的宽度；比例大于 1，图宽大于图高；比例等于 1，图形为高宽相等的正方形。一旦图形生成，在 SPSS 中其长宽比就不能改变了。

3. 【当前设置】

在该栏中设置生成图形的参数。

- (1) 【字体】。设置图形中的文字字体，单击向下箭头，在下拉列表中选择一种字体。默认的是没有修饰的普通字体。SPSS 20.0 版软件可以设置中文字体，在下拉列表中可以找到。

(2) 【样式循环设置】。在其中设置新生成图形的填充方式。下拉列表中的选项有：

- 【仅在颜色之间循环】。用不同颜色区别图形不同的分类，不使用底纹图案。
- 【仅在图案之间循环】。只用不同底纹区别图形不同的分类，而不使用颜色。如果显示器为单色显示器，选择此项可以在屏幕上获得比较好的图形显示效果。

(3) 【框架】。图形框设置栏，本栏内有两个复选项。

- 【外部】。外框，即在整个统计图(包含标题和图例等)的外围加框。
- 【内部】。只对统计图形加边框。

(4) 【网格线】。提供两种坐标轴格线。

- 【刻度轴】。显示刻度坐标轴格线。
- 【类别轴】。显示分类坐标轴格线。

#### 4. 【样式循环】

该栏设置图形外观样式参数。

单击【颜色】、【线】、【标记】和【填充】按钮打开相应的对话框，设置图形的颜色、线条、标记和填充的样式。

(1) 设置图形颜色

单击【颜色】按钮，打开如图 1-23 所示的【数据元素颜色】对话框。左面是【要编辑的样式】选择项装饰编辑栏。

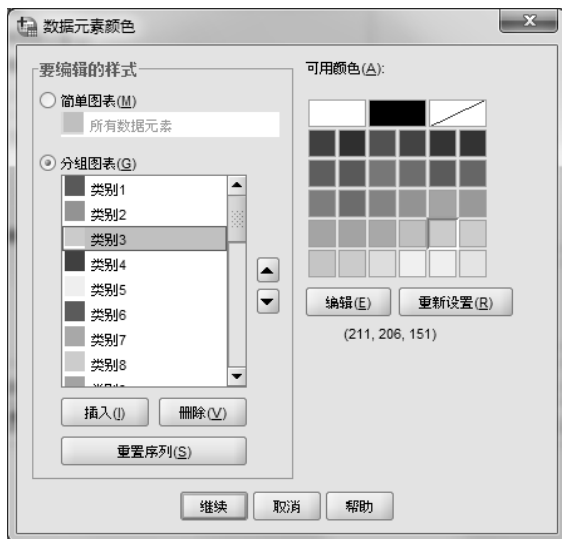


图 1-23 【数据元素颜色】对话框

① 【简单图表】。是默认的，简单图形用单一颜色标注所有图形元素。默认颜色是淡黄色，显示在该项目下方。例如，做出的柱形图每个柱都是淡黄色。要改变默认颜色，只要在右面的调色板中选择一种合适的颜色，单击之即可。

② 【分组图表】。对于一组图形需要有不同颜色表示时，可选择此项。图形颜色的选取可按下面栏中自上而下的顺序选择。

- 如果要改变颜色使用顺序，只要单击选择的一种颜色，然后单击向上或向下箭头按钮，就可以向前、向后移动使用顺序。
- 如果要改变栏中的某一种颜色，只要单击这个颜色块，然后到右面的调色板中选择一种合适的颜色，单击之即可。
- 可以在【可用颜色】的调色盘中选择一种颜色，单击【插入】按钮，将其插入到颜色列表中；对不想使用的颜色，在列表中选中后单击【删除】按钮，将其放回【可用颜色】列表中。

③ 编辑调色板有 3 种方式。单击调色板下方的【编辑】按钮，打开如图 1-24(a) 所示的【选择颜色】对话框，3 个选项卡是选择合适颜色的 3 种方式。

- 【样品】选项卡。通过样品块方式选择颜色，只要鼠标单击选中的颜色，就有一小块样品显示在右侧【最近】的格子中。在下面的【预览】栏中察看各种符号颜色是否合适。
- 【HSB】色相、饱和度、明度方式。方块中是在样品块方式中选择的颜色，只不过饱和度和明度是渐变的。只要用鼠标拖曳其中的圆圈到饱和度和明度合适位置松开鼠标即可。是否合适，察看下面的预览窗口中的各种符号。
- 【RGB】红绿蓝基本色参数方式。熟悉参数的读者可以使用这种方式。

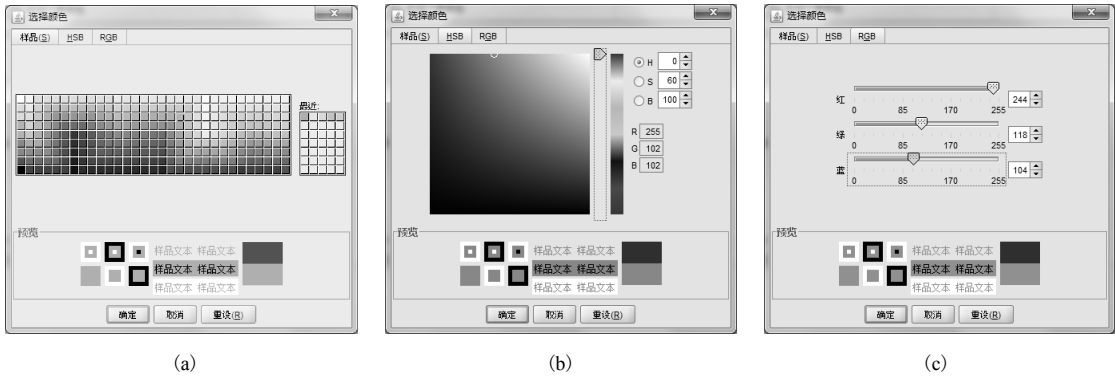


图 1-24 3 种方式改变调色板颜色的对话框

一般凭直觉调整颜色的读者只要选用前两种方式就可以满足要求。设置颜色完成后单击【确定】按钮返回【数据元素颜色】对话框。

单击【继续】按钮返回主对话框。

(2) 设置线型

单击【线】按钮打开【数据元素线】对话框，如图 1-25 所示。

在【要编辑的样式】栏内有两种目标供选择：

①【简单图表】。对所有数据元素都使用一种线型。默认线型是细直线。样品显示在选择项下方。若要改变这个基本线型，只要在右面的线型列表中选择一种，单击之即可。例如只有一根折线的折线图，可以选择这种方式。

②【分组图表】。例如由两条以上的直线或折线组成的图形，需要两种以上线型，以便区别，则可以选择此项。

- 改变线型使用顺序，线型使用顺序按线型列表自上而下的顺序。要改变顺序，只要选择一种线型后单击之，然后单击向上或向下箭头按钮移动该线型的位置即可。
- 也可以用【可用线条】表中选中的线型。单击【插入】按钮插入到列表中所选线型的下方。对不使用的线型，可以在列表中选择后，单击【删除】按钮将其送回【可用线条】表中。
- 改变列表中某线型，只要单击该线型，然后在右面可选择的线型中选一种，单击之即可。
- 若要恢复默认状态，只需要单击【重置序列】按钮。

(3) 设置标记

单击图 1-22 中的【标记】按钮，打开【数据元素标记】对话框，见图 1-26(a)。设置图形中数据点所用标记。

(4) 设置图案

单击图1-22 中的【填充】按钮，打开【数据元素填充】对话框，图1-26(b)为设置图案对话框。设置有框图形，如柱形图、饼图等，填充内部使用的图案或称底纹。操作方法与设置线型相同。



图 1-25 【数据元素线】对话框

以上所有在【图表】选项卡中的设置，在单击【确定】按钮后，只对设置后产生的图形生效。



(a) 【数据元素标记】对话框



(b) 【数据元素填充】对话框

图 1-26 设置标记和图案的对话框

1.3.8 输出表格参数设置

【枢轴表】选项卡用来设置默认的输出表格样式及有关参数，如图 1-27 所示。在该选项卡中，可对新的表格输出设置外观。



图 1-27 【枢轴表】选项卡

在【表格外观】栏中选择一个表格外观样式，被选择的表格样式显示在【样本】栏中。单击【应用】或【确定】按钮，新表格按选择的形式生成。【表格外观】中的表格样式文件保存

在 SPSS 系统所安装位置的一个文件夹中。选中的表格样式文件的保存位置和文件名，显示在【表格外观】栏第一行。默认的外观是外框粗实线，内部细实线，不加任何修饰的普通表格。

我们还可以使用 SPSS 提供的【表格外观】功能建立表格样式，即双击输出窗口中的表格，在表格编辑窗口中选择【格式】菜单的【表格外观】命令，打开【表格外观】对话框，选择一种基本样式。再单击【编辑外观】按钮打开【表格属性】对话框，改变表格各部位的参数以建立自己的表格样式。详见 3.2.3 节。

1.3.9 文件默认存取位置设置

【文件位置】选项卡中设置 SPSS 启动后打开和保存文件的位置。系统默认的位置显示在每个选择项的编辑栏中。一般都是 Windows 用户的 My Document 文件夹。可以单击【浏览】按钮对文件位置进行重新设置。此功能为减少启动 SPSS 后查找数据文件或其他类型文件的操作而设置。

【文件位置】选项卡如图 1-28 所示。



图 1-28 【文件位置】选项卡

(1)【打开和保存对话框的启动文件夹】。栏中设定打开和保存对话框所使用的默认文件夹。  
注意：指定的文件夹必须事先建好。如果指定了一个不存在的文件夹，单击【确定】按钮后，系统会给出警告信息，并要求改变。

①【指定文件夹】。在该栏中指定保存数据文件的文件夹和保存其他文件的文件夹。

- 在【数据文件】栏直接输入或单击【浏览】按钮定位，指定数据文件位置。即单击【浏览】按钮，打开【默认数据文件夹】对话框，改变系统默认的设置。在【查找范围】下

拉列表中确定文件夹位置，在【文件夹名称】框中输入文件夹的名字，单击【设置】按钮完成设置。

- 在【其他文件】栏直接输入或单击【浏览】按钮定位，指定非数据文件位置。

②【最后使用的文件夹】，选择此项，启动 SPSS 后，打开或保存操作直接使用上次从 SPSS 退出时最后使用的文件夹。

(2)【会话日志】。在该栏中指定 SPSS 运行时产生的日志文件自动保存的位置和形式。

①【日志中的记录语法】。选择此项，每次运行会把语句写进日志文件。系统默认选择此项，习惯于编程的人员更要选择此项。日志中记录的语句既包括写在语句窗中的程序，也包括对话框操作时调用的命令、设置的参数等形成的语句程序，对下一次修改程序减少程序输入量很有用。

② 日志文件续写方式设定：

- 【附加】。每次运行的语句接在前一次运行语句记录后面，存入日志文件。
- 【覆盖】。每次运行语句存入日志文件时覆盖前一次存入的内容。

③ 设定日志文件名及存储路径。在【日志文件】后面直接输入或单击【浏览】按钮，展开保存日志文件的【另存为】对话框，指定保存日志文件的存储位置和文件名。

(3)【临时目录】。临时文件路径设置。读者可以直接输入或者单击【浏览】按钮，打开【临时文件夹】对话框，设置在统计处理过程中的临时文件的存放位置和文件名。临时文件往往需要较大空间，例如 200MB 的数据文件，需要大于 400MB 的临时文件空间。不用的临时文件及时删除。

(4) Number of Recently Used Files to List(最近使用过的文件数设定)。该栏设定最近使用文件数目。它控制显示在【文件】菜单的【最近使用的文件】中的文件名数目。改变参数框中的数字，即可达到目的。

### 1.3.10 缺失值处理

数据收集过程中会由于种种原因，使得数据在一定程度上缺失，或大量缺失。为弥补缺失值带来的信息损失和解决分析中的问题，会采用各种插值方法，给缺失值较为合理的值，以得到完整的数据集。

【多重归因】选项卡设置归因数据的标志，以区分原始数据；确定输出内容，以便研究人员确定归因数据的影响。

【多重归因】选项卡如图 1-29 所示。

(1)【归因数据标记】栏设置数据集中插值得到的原缺失值数据显示方式。在【单元格背景色】栏中单击向下箭头，在调色盘中选择一种颜色；在【字体】下拉菜单中选择一种字体，单击右侧的【B】图标，可选择或者取消对字体加粗。

(2)【分析输出】栏确定与插值后数据处理结果的输出内容。

- ①【观测值与归因数据结果】输出观测值与归因数据一起分析的结果。
- ②【仅观测值结果】仅输出对观测值分析的结果。
- ③【仅归因数据结果】仅输出对归因值分析的结果。

以上选择项只能择其一。还有两个复选项是系统默认选项：

- ①【汇聚结果】。
- ②【诊断统计结果】。



图 1-29 【多重归因】选项卡

## 1.4 统计分析功能概述

SPSS 20.0 的统计分析功能主要集中在以下 3 个方面。

### 1. 统计分析函数

统计分析函数共 18 类 195 个函数，如算术函数、CDF 与非中心 CDF 函数、转换函数、当前日期时间函数、日期算法函数、日期生成函数、日期提取函数、反 DF 函数、PDF 与非中心 PDF 函数、随机数函数、查找函数、显著性函数、统计函数、字符函数、时间间隔生成函数、时间间隔提取函数和缺失值函数等。

### 2. 统计分析过程

在【分析】菜单中有 22 类，共 73 个分析过程，另外还有可以使用语句实现分析而没有收入窗口化的 SPSS 软件中的统计方法。在窗口化软件中的方法都可以使用编程语句实现。

### 3. 统计图

统计图可以直观表达数据特征和统计分析的结果。大致可以分为以下两类。

① 在【图形】菜单中包括条形图、线图、面积图、圆图、高低图、帕累托图、控制图、箱图、误差条图、散点图、直方图、P-P 概率图、Q-Q 概率图、序列图等，并有一套灵活、完整的对统计图进行编辑的方法。这些统计图是对数据统计特征的描述，可以作为初步的统计分析和对数据特征的认识。

② 绝大多数统计分析方法都能产生统计图。这些图有些可以通过【图形】菜单中的图形功能



产生,有的则直观表达统计分析的结果,一般都在分析过程对话框的二级窗口【绘制】对话框的选项中。

**注意:**各种统计分析方法使用的条件,正确选择和充分利用 SPSS 中的各种统计分析功能,辛辛苦苦获得的数据通过定量分析,一定能挖掘出有用的信息。

## 1.5 数据与变量

### 1.5.1 常量与变量

#### 1. SPSS 常量

常用的 SPSS 常量有数值型、字符型、日期型和日期时间型。

① 数值型常量就是 SPSS 语句中的数字。一般使用两种书写方式:一种是普通书写方式,例如 26、38.4 等;另一种是科学计数法,即用指数表示数值的计算机书写方式,用于表示特别大或特别小的数字,例如 1.23E18(或 1.23E+18)表示  $1.23 \times 10^{18}$ , 2.56E-16 表示  $2.56 \times 10^{-16}$ 。

② 字符串常量是被单引号或双引号括起来的一串字符。如果字符串中带有“'”字符,则该字符串常量必须使用双引号括起来,例如“BOY'S BOOK”。在数据窗口中的字符串不使用引号。

③ 日期型常量表示方法很多,可以使用表 1-1 中所列的各种格式。

#### 2. SPSS 变量及其属性

SPSS 中的变量除应定义【变量名】(SPSS 汉化为【名称】)外还应该定义 4 个属性:变量【类型】(type);格式——变量【宽度】(width)、【小数位数】(decimal);【缺失值】定义(missing value);【测度类型】(measure) (SPSS 汉化为【度量标准】)。另外,为输出查看方便还可以定义变量【标签】(label)和【值】标签(values);变量在数据窗口中的显示【列】宽度(columns)、【对齐方式】(align);SPSS 的变量至少要定义变量名和变量类型,其他属性可以采用默认值。

(1) 【变量名】命名应该遵循的原则

① SPSS 的【变量名】最多可长达 64 字节,相当于 64 个英文字符或 32 个汉字的长度。

② 首字符不能是数字,必须字母打头,其后可为除“?”、“-”、“!”、“\*”、“#”、“\$”和空格以外的字符或数字。但应该注意,不能以下划线“\_”和圆点“.”作为自定义变量名的最后一个字符。

③ 【变量名】不能与 SPSS 保留字相同。SPSS 的保留字包括 ALL、AND、BY、EQ、GE、GT、LE、LT、NE、NOT、OR、TO、WITH。

④ 系统不区分变量名中的大小写字符,例如,ABC 和 abc 被认为是同一个变量。

(2) 变量类型与默认长度

SPSS 变量有 3 种基本类型:数值型、字符型、日期型(或日期时间型)。数值型变量又按不同要求分为 5 种。系统默认的变量类型为标准数值型变量(numeric)。每种类型的变量由系统给定默认长度。小数点或其他分界符包括在总长度之内。变量的系统默认长度可以用【编辑】菜单中的【选项】命令重新设置。

① 标准数值型变量(numeric),默认总长度为 8,小数位数为 2。标准数值型变量的值用标

准数值格式显示，小数点用圆点，可以用标准数值格式输入，也可以用科学计数法输入。使用科学计数法输入时，显示出来的还是标准数值格式的数值。

② 带逗点的数值型变量(**comma**)，默认总长度为 8，小数位数为 2。其值在显示时整数部分自右向左每三位用一个逗点作分隔符，用圆点作小数点。定义为此格式的变量，在输入时可以带逗点，也可以不带逗点，还可以用科学计数法输入。使用科学计数法输入时，显示的还是用圆点作小数点，逗点作三位分隔符的数值。

③ 圆点数值型变量(**dot**)，默认总长度为 8，小数位数为 2。显示方式与带逗点的数值型变量正好相反。整数部分自右向左每三位用一个圆点作分隔符，用逗点作小数与整数间的分界符。输入时可以带圆点，也可以不带圆点。还可以用科学计数法输入。

④ 科学计数法(**scientific notation**)，默认总长度为 8，小数位数为 2。  
数值很大或很小的变量可以使用科学计数法，这种变量的值可以有指数部分，也可以没有指数部分。表示指数的字母可以用 E，也可以用 D。指数部分可以带正负号，正号可以省略，甚至指数部分不用字母 D 或 E，只用符号表示也是可以接受的。例如，表示一百二十三，可以用以下方式输入：1.23E2、123、1.23D2、1.23E+2、1.23+2 等。

⑤ 带美元符号的数值型变量(**dollar**)，默认总长度为 8，小数位数为 2。其值在显示时有效数字前带有“\$”，变量总长度包括“\$”符号在内，其余规定与标准数值格式相同。输入数据时可以带“\$”，也可以不带。显示在数据表格中的数值由系统自动加上“\$”符号和分隔符。可以用科学计数法输入，如果数值不超过定义的长度，则显示在数据表格中的数值自动变换为定义的格式。

带美元符号的数值型变量的具体格式还可以从格式列表框中选择，见表 1-1。

表 1-1 带美元符号的数值型变量格式列表框选项举例

格 式	总长度	小数位数	格 式	总长度	小数位数
\$##	3	0	\$###.##	7	2
\$###.##	6	0	\$###.##	9	2

选定的格式只对数据表格中的显示形式有效，当输入的数值小数位数超过格式规定时，系统自动进行四舍五入处理。如果输入的整数位数超出规定的格式，则显示时自动去掉作为三位分隔符的逗号。

读者应该根据数据中最大数值的位数指定显示格式，以便使显示与输入的值一致。  
⑥ 自定义型(**custom currency**)是一种由读者用【编辑】菜单的【选项】功能来定义的，定义方法参见 1.3 节。

⑦ 日期型变量(**date**)  
SPSS 的日期型变量可以表示日期，也可以表示时间。日期型变量的值按指定的格式输入和显示，不能直接参与运算。要使用日期型变量的值进行运算，必须通过有关的日期函数转换，详见第 5 章。

(3) 变量格式  
对数据的宽度(**width**)和小数位数(**decimal**)的要求。对数值型变量要定义宽度和小数位数。对字符型变量只定义宽度。日期型一般使用默认宽度，一旦日期格式确定了，宽度就确定了，不用再进行设置。

(4) 变量的标签与值标签  
① 变量标签(**variable labels**)是对变量名附加的进一步说明。变量名只能由不超过 64 个

字符组成，如果 64 个字符不足以表明变量的含义，或变量比较多时，则需要用变量标签对变量名的含义加以解释。如果 SPSS 运行在中文环境下，也可以给变量名附加中文标签，见表 1-2。

② 变量值标签(values)是对变量可能取值附加的进一步说明。对分类变量往往要定义其取值的标签。如果 SPSS 运行在简体中文版的 Windows 环境下，也可以给变量值附加中文标签，见表 1-2。

变量标签和变量值标签是可选择的属性，可定义，也可不定义。为了对输出信息进行解释并得出结论，建议使用中文标签。在输出窗口的输出表格中可以使用标签表明变量和变量值，这就要通过【编辑】菜单中的【选项】进行设置。

(5) 变量的显示格式

① 宽度(columns)显示数据的宽度。应该区分定义变量类型时指定的宽度与定义显示格式时的宽度。显示宽度应该综合考虑变量类型定义的总长度和变量名所占宽度。显示宽度不影响机内值，不影响分析运算结果，只影响显示。

② 对齐方式：分为左对齐、右对齐、中间对齐。一般情况下，数值型变量默认右对齐；字符型变量默认左对齐，也可以指定为中间对齐方式。

(6) 缺失值(missing)

已经输入的失真数据、没有测到或没有记录的数据，以特殊的数字或符号输入到数据文件中，统称为“缺失值”，都应该加以定义。在分析时不能使用，或需要单独处理。在 SPSS 中，字符型变量默认的缺失值为空格；数值型变量的缺失值没有默认值，需要定义。各分析过程对缺失值的处理都有默认的方法，也可以由读者指定选择项，定义如何处理这些缺失值。

(7) 变量测度方式

变量测度方式是指变量是如何测量的。

① 等间隔测度变量，即按与尺度的比例测度的变量，也可称为尺度变量(scale)，如身高、体重。

② 有序变量(ordinal)，如表示职称、职务、对某事物的赞同程度的变量，是分类变量中有顺序特性的一种，可以用有序的数字作为代码。设置了值标签的变量被认为是有序的分类变量，可以作为分组变量，也可以参与某些分析过程的运算。

③ 名义变量(nominal)，是无序的分类变量，取值是无法度量的。只能作为分组变量使用。如表示民族、宗教信仰、党派等的变量。

分类变量值为数值时，它与尺度变量的分界默认值为 24。当变量的独立数值的个数大于 24 时被认为是尺度变量，小于 24 时被认为是有序的分类变量。这个数值也可以通过【编辑】菜单的【选项】重新设定。

表 1-2 变量和变量值标签

变 量	变 量 标 签	变 量 值	变量值标签
Gender	性别	f	男
		m	女
Height	身高	1	<=1.49m
		2	1.50~1.59m
		3	1.60~1.69m
		4	1.70~1.79m
		5	>=1.80m

1.5.2 操作符与表达式

SPSS 的基本运算共有 3 种：数学运算、关系运算、逻辑运算，运算符见表 1-3。

表 1-3 SPSS 的基本运算符

数学运算操作符	关系运算符	逻辑运算符
+: 加	< (LT): 小于	& (and): 与
-: 减	> (GT): 大于	(or): 或
*: 乘	<= (LE): 小于等于	~ (not): 非
/: 除	>= (GE): 大于等于	
** : 幂	= (EQ): 等于	
() : 括号	~= (NT): 不等于	

1. 算术运算符与算术表达式

算术运算符可以连接数值型的常量、变量和函数来构成算术表达式，其运算结果为数值型常量。例如， $X+Y**2/(A+B)-1+ABS(A*Z)$  就是一个合法的算术表达式。在算术运算表达式中，运算的优先顺序：括号、函数、乘方(幂)、乘或除、加或减的顺序，同一优先级的位于左面的先算。

2. 比较算符与比较表达式

比较算符建立的是两个量之间的比较关系，由系统判断关系是否成立。如果比较关系成立，则比较表达式的值为逻辑值“真”，否则为“假”。相互比较的两个量必须类型一致。无论进行比较的两个量是字符型还是数值型，比较的结果均是逻辑型常量。比较算符表中列出的比较算符均有两种表示方法。表 1-3 括号中的比较算符与括号前的算符是等价的。例如， $A>3$  和  $A\text{ GT }3$  是等价的。如果  $A=4$ ，则表达式  $A>3$  的值为真，其值为 1；如果  $A=3$ ，则表达式  $A>3$  的值为假，其值为 0。

3. 逻辑运算符与逻辑表达式

逻辑运算符即布尔运算符。表 1-3 中，括号前的运算符与括号中的运算符等价，例如， $A\&B$  与  $A\text{ and }B$  是等价的。逻辑运算符与逻辑型的变量或其值构成逻辑表达式。逻辑表达式的值为逻辑型常量。

(1) 与运算。 $\&$ (或 and)前后的两个量均为真时，逻辑表达式的值为“真”，否则为“假”。例如，逻辑表达式  $A>B\&C>0$ ，当  $A$  的值大于  $B$  的值且  $C$  为正数时，该逻辑表达式的值为“真”。如果  $A=3$ ， $B=2$ ， $C=-6$ ，则该逻辑表达式的值为“假”。

(2) 或运算。 $|$ (或 or)前后的两个量只要有一个为“真”时，逻辑表达式的值即为“真”。只有当运算符前后两个量均为假时，逻辑表达式的值才为“假”。

例如，逻辑表达式  $A>B|C>0$ ，只要  $A>B$  成立，无论  $C$  为何值，表达式的值均为“真”。或者只要  $C>0$  成立，无论  $A$  与  $B$  为何值，该表达式的值也为“真”。当  $A<B$ ，同时  $C\leq 0$  时，该逻辑表达式的值为“假”。

(3) 非运算。 $\sim$ (或 NOT)是前置运算符，它对其后面的量作非运算。NOT 后面的量值为“真”，则 NOT 运算结果为“假”；后面的量值为“假”，则 NOT 运算的结果为“真”。

例如，逻辑表达式  $\text{NOT}(A>0)$ ， $A$  为正数，则逻辑表达式的值为“假”； $A$  为负数或  $A$  为 0，则逻辑表达式的值均为“真”。

1.5.3 观测

在【数据编辑器】的【数据视图】窗口中是个二维表格。每行都是数据文件的一个记录，

在统计学中称作“一个观测”，在 SPSS 的菜单中或帮助信息中用“个案”这个词表示。每个个案由各变量的一定的值组成，是对一个事件，或者说是由一个被观测对象的各种特征的实测值或派生值组成的，因此，相对“变量”来说可以称之为“观测”。单元格中的数值既是某个变量值，也是某个观测中的一个值，因此可以称之为××变量值，也可以称之为某个观测的某个变量值。

1.5.4 SPSS 函数

SPSS 有 18 类函数，见表 1-4。函数的表示方法是在函数关键字后面括号中写入函数自变量。函数自变量有的要求使用单个值或变量名，有的要求使用“:”隔开的多个变量名，还有允许使用表达式。当然，如果使用变量名或带有变量名的表达式作为自变量，则必须在使用该函数之前对这些变量赋值。下面列出 SPSS 函数，函数类型即函数值的类型。

函数中使用的符号说明：*numexpr* 表示数值型表达式；*radians* 表示以弧度为单位的角度。

1. 算术函数 (Arithmetic)

算术函数共 13 个。

(1) ABS (*numexpr*) 数值型函数。函数值为数值表达式的绝对值。

表 1-4 SPSS 函数类型清单

序 号	类 型		数 量
1	Arithmetic	算术函数	13
2	CDF & Noncentra CDF	累积分布函数	30
3	Conversion	转换函数	2
4	Current Date/Time	当前日期、时间函数	4
5	Date Arithmetic	日期算术函数	3
6	Date Creation	日期生成函数	6
7	Date Extraction	日期提取函数	11
8	Inverse DF	反分布函数	18
9	Miscellaneous	混杂函数	4
10	Missing Values	缺失值函数	6
11	PDF & Noncentra PDF	概率密度函数	27
12	Random Number	随机数函数	22
13	Search	查找函数	10
14	Significance	显著性函数	2
15	Statistical	统计函数	7
16	String	字符函数	26
17	Time Duration Creation	时间间隔生成函数	4
18	Time Duration Extraction	时间间隔提取函数	8

(2) ARSIN (*numexpr*) 数值型函数。函数值为数值表达式的反正弦值，单位为弧度，自变量 *numexpr* 范围在-1~1 之间。

(3) ARTAN (*numexpr*) 数值型函数。函数值为数值型自变量表达式 *numexpr* 的反正切值，单位为弧度。

(4) COS (*radians*) 数值型函数。函数值为单位为弧度的自变量表达式 *radians* 的余弦值。

(5) EXP (*numexpr*) 数值型函数。函数值为以 e 为底, 以括号中的自变量表达式 *numexpr* 为指数的幂值。应该注意, 若指数太大或函数值太大, 其结果会超出 SPSS 的计算范围。

(6) LN (*numexpr*) 数值型函数。函数值为以 e 为底的自然对数值, 自变量数值表达式 *numexpr* 必须是数值型, 而且要大于 0。

(7) LNGAMMA (*numexpr*) 数值型函数。函数值为数值表达式 *numexpr* 的完全 Gamma 函数的对数。表达式必须是数值型的, 且其值必须大于 0。

(8) LG10 (*numexpr*) 数值型函数。函数值为以 10 为底的对数值, 数值表达式 *numexpr* 必须是数值型, 而且值要大于 0。

(9) MOD (*numexpr*, *modulus*) 数值型函数。函数值为数值表达式 *numexpr* 除以模数 *modulus* 的余数。两个自变量必须是数值型, 模数不能为 0。

(10) RND (*numexpr*, [*mult*, *fuzzbils*]) 数值型函数。函数值为数值表达式 *numexpr* 的值取四舍五入后的整数。第 2、3 个参数是可选项。因此有 3 种情况, 前两种是常用的:

① RND (*numexpr*)。如果 RND 函数只有一个数值型参数 *numexpr*, 则函数值是最接近该参数值的整数。尾数为 0.5 的数, 其该函数值为舍五进一。例如 RND(-7.5)=-8。

② RND (*numexpr*, *mult*)。两个参数的 RND 函数, 第 2 个参数必须是不为 0 的数值型变量或数值, 默认为 1。函数值是将 *numexpr*, 四舍五入成 *mult* 值的整数倍。例如 RND(4.55,0.2)=4.6。在转换菜单选择计算变量功能建立新变量时, 选择一个参数, 指定 RND1, 使用两个参数时, 指定 RND2……

(11) SIN (*radians*) 数值型函数。自变量 *radians* 是以弧度为单位的角度, 函数值为弧度角的正弦值。

(12) SQRT (*numexpr*) 数值型函数。函数值为一个正数的平方根。数值表达式 *numexpr* 的值必须大于等于 0。

(13) TRUNC (*numexpr*, [*mult*, *fuzzbils*]) 数值型函数。函数值为数值表达式 *numexpr* 的值被截去小数部分的整数。第 2、3 个参数是可选的。因此有 3 种情况, 前 2 种是常用的。

## 2. 累积分布函数(CDF & Noncentral CDF)

累积分布函数共 30 个, 详见第 4 章。

## 3. 转换函数(Conversion)

转换函数共两个。

(1) NUMBER (*strexp*, *format*) 数值型函数。当字符串内容为一串数字时, 该函数返回字符串表达式作为数字的值, 返回的函数值可以参与运算。第 2 个表达式为一个数值格式, 用来读取字符串表达式中的数字。

如果 *name* 是一个由 8 个数字组成的字符串, NUMBER (*name*, *f8*) 就是由这些数字表示的数值。如果字符串不能使用指定的格式, 该函数返回系统缺失值。

(2) NUMBER (*stringDate*, *Date11*) 数值型函数。把内容为标准格式(dd-mm-yyyy)日期的字符串转换成描述该日期的秒数。如果字符串不能使用标准格式读取, 则函数值是系统缺失值。第一个自变量是字符型, 自变量的值为与 *Date11* 格式相应的日期。

如果我们定义了字符串格式的自变量, 输入了与 dd-mm-yyyy 相应的日期, 可以使用该函数将字符串变量转换为日期变量。

(3) `STRING (numexpr, format)` 字符型函数。根据 `format` 所设定的格式将数值表达式转换为字符串。例如, `string(-1.5, F5.2)` 返回字符串 ‘-1.50’。第 2 个自变量 `format` 必须是一个数值的格式。

注意: 数值与数字有区别, 以上所讲的数值是数, 数字指的是表现为数字的字符。

#### 4. 当前日期/时间函数 (Current Date/Time)

当前日期/时间函数共 4 个。

#### 5. 日期算术函数 (Date Arithmetic)

日期算术函数共 3 个。

#### 6. 日期生成函数 (Date Creation)

日期生成函数共 6 个。

#### 7. 日期提取函数 (Date Extraction)

日期提取函数共 11 个, 有关日期的函数和应用见第 4 章。

#### 8. 反分布函数 (Inverse DF)

反分布函数共 18 个, 详见第 4 章。

#### 9. 混杂函数 (Miscellaneous)

混杂函数共 4 个。

(1) `$Casenum` 无参数函数。其值为当前观测(或称个案)的顺序号。对每个观测, `$Casenum` 是读取的并包括这个观测的观测号, 格式是 F8.0。`$Casenum` 的值不一定是数据编辑窗中的行号, 如果文件排序或者新的观测代替了文件末尾之前的观测, 这个值也会改变。

(2) `LAG(variable)` 数值型或字符型函数。函数值是前一个观测的变量值。

(3) `LAG(variable[, n])` 数值型或字符型函数。函数值是前一个或前  $n$  个观测的变量值。第 2 个自变量是可选的。 $n$  必须是正整数, 默认值为 1。例如 `prev4=LAG(gnp, 4)` 的值为当前观测之前的第 4 个观测的变量 `gnp` 的值。

(4) `VALUELABEL(varname)` 字符型函数。函数值是变量值的标签, 当该值没有标签时函数值是空字符串。自变量 `varname` 必须是变量名, 不能是表达式。

#### 10. 缺失值函数 (Missing Values)

缺失值函数共 6 个。

(1) `$SYSMIS` 数值型函数, 产生系统缺失值。常用于判断并记录缺失值。例如, 如果取得的数据中有小于 1.4 m 的观测, 而身高 < 1.4 m 就不能参与一项研究。可以执行语句:

```
IF (height<1.40) height=$Sysmis.  
EXECUTE.
```

就可将身高变量值小于 1.4 的身高值改为圆点。可以在【转换→计算变量】, 打开相应对话框完成操作。

(2) `MISSING(variable)` 逻辑型函数。如果变量具有缺失值, 则返回 1 或者 True。自变量应该是当前工作数据文件中的变量名。

(3) NMISS (*variable* [,...]) 数值型函数。函数值是自变量表中各自变量具有的系统缺失值或用户缺失值的总数。此函数需要至少一个自变量, 这些自变量必须是当前工作数据文件中的变量名。

(4) NVALID(*variable*[,...]) 数值型函数。函数值为自变量表中的变量具有的合法非缺失值的总数。函数要求至少有一个自变量, 自变量应该是当前工作数据文件中的变量名。

(5) SYSMIS (*numvar*) 逻辑型函数。如果 *numvar* 的值为系统缺失值, 则函数值为 1 或者 true。自变量 *numvar* 必须是工作数据文件中的一个数值型变量的变量名。

(6) VALUE (*variable*) 数值型或字符型函数。忽略用户定义的缺失值, 返回变量值。自变量必须是工作数据文件中的变量名。

需要说明的是, 函数和简单的算术表达式用不同的方法处理缺失值。

① 在表达式  $(var1+var2+var3)/3$  中, 如果一个观测的 3 个变量中任意一个是缺失值, 则运算结果就是缺失值。

② 在表达式 MEAN(*var1, var2, var3*) 中, 仅当一个观测的所有变量的值都是缺失值时, 运算结果才是缺失值。

③ 对于统计函数, 可以在函数名后面, 指定非缺失值的最小数。为此, 在函数名后面打一个半角圆点, 以及至少要有的非缺失值数目, 例如 MEAN.2(*var1, var2, var3*)。

## 11. 概率密度函数(PDF& Noncentral PDF)

概率密度函数共 27 个, 详见第 4 章。

## 12. 查找函数(Search)

查找函数共 10 个(与其他类拆分的有 8 个)。

(1) ANY (*test, value* [, *value*...]) 逻辑型函数。如果 *test* 的值与其后的 *value* [, *value*,...] 中的某一数值匹配, 那么函数值为真, 返回 1 或 True; 否则, 函数值为假, 返回 0 或者 False。该函数要求至少有两个自变量。例如 ANY(*var1, 1, 3, 5*), 如果 *var1* 的值是 1 或 3 或 5, 则函数值为 1; 若 *var1* 为其他值, 则函数值为 0。该函数还可以用来在变量表或表达式表中扫描一个值。例如 ANY(1, *var1, var2, var3*), 如果在 3 个指定的变量中任意一个变量有 1 值, 则函数值为 1; 若所有变量的值都不是 1, 则函数值为 0。

(2) RANGE (*test, lo, hi* [, *lo, hi*,...]) 逻辑型函数。如果 *test* 的值包含在由 *lo, hi* 所定义的范围内, 则函数值为 1 或者 True; 否则为 0 或者 False。所有变量必须都为数值型或都为字符型, 并且所设置的 *lo, hi* 变量的大小顺序必须为  $lo \leq hi$ 。注意, 不同地区使用不同语言, 对自变量为字符型的情况, 同一个函数运算结果可能有很大区别。本函数按 ASCII 码顺序运算。

另有 6 个字符串函数: CHAR.INDEX(2)、CHAR.INDEX(3)、CHAR.RINDEX(2)、CHAR.RINDEX(3)、REPLACE(3)、REPLACE(4)。重复出现在字符串函数类中。

另外, SPSS 20.0 把 Max、Min、Range 也列入了查找函数。前面两个在统计函数中重复出现, 这里不再解释。

## 13. 显著性函数(Significance)

显著性函数共两个。



(1) SIG.CHISQ(*quant*, *df*) 数值型函数。其值为自由度为 *df* 的卡方分布中的值大于 *quant* 的累积概率。

(2) SIG.F(*quant*, *df1*, *df2*) 数值型函数。其值为自由度是 *df1*、*df2* 的 F 分布中值大于 *quant* 的累积概率。

## 14. 统计函数 (Statistical)

统计函数共 7 个。

(1) CFVAR (*numexpr*, *numexpr*[,...]) 数值型函数。函数值为自变量 (或数值表达式 *numexpr* 的值) 的变异系数 (标准差除以均值)。此函数要求有两个或两个以上自变量。自变量必须为数值型, 而且必须有合法值。

(2) MAX (*value*, *value*[,...]) 数值型函数或字符型函数。函数值为自变量 *value* 所有合法值的最大值。至少需要两个以上 *value*。

(3) MEAN (*numexpr*, *numexpr*[,...]) 数值型函数。函数值为多个数值表达式 *numexpr* 的算术平均数。数值表达式至少需要两个以上。

(4) MEDIAN (*numexpr*, *numexpr*[,...]) 数值型函数。函数值为多个数值表达式值 *numexpr* 的中位数。至少需要两个以上数值表达式。

(5) MIN (*value*, *value*[,...]) 数值型函数或字符型函数。函数值为具有合法值的自变量 *value* 的最小值。至少需要两个以上 *value*。

(6) SD (*numexpr*, *numexpr*[,...]) 数值型函数。函数值为所有数值表达式的标准差。这个函数需要两个或两个以上自变量, 自变量可以是表达式, 也可以是非缺失的合法值, 而且必须为数值型。

(7) SUM (*numexpr*, *numexpr*[,...]) 数值型函数。函数值为所有数值表达式值的累加和。这个函数需要两个或两个以上非缺失合法值。自变量可以是数值、数值型表达式。

(8) VARIANCE (*numexpr*, *numexpr*[,...]) 数值型函数。函数值为所有数值表达式的方差。这个函数需要两个或两个以上自变量。自变量可以是表达式, 但必须是数值型。

## 15. 字符串函数 (String)

字符串函数共 26 个。

(1) CHAR.INDEX (*haystack*, *needle*) 数值型函数。返回一个整数, 它表明 *needle* 代表的字符串在 *haystack* 代表的字符串中第一次出现的起始位置。如果返回值为 0, 表明字符串 *needle* 在字符串 *haystack* 中不存在。在函数表中, CHAR.INDEX(*var1*, 'abcd') 将返回整个字符串 “abcd” 在字符串变量 *var1* 的起始位置。函数列表中该函数的函数名为 CHAR.INDEX (2), 意为两个自变量。

(2) CHAR.INDEX (*haystack*, *needle*, *divisor*) 数值型函数。见前一个函数。其第 3 个自变量 *divisor* 是可选择的, 它必须是一个整数, 表明将字符串 *needle* 均匀地分为要查询的独立的子字符串的字符数。例如, CHAR.INDEX(*var1*, 'abcd', 1) 返回字符串中任意一个字符在字符串变量 *var1* 代表的字符串中第一次出现的位置。CHAR.INDEX(*var1*, 'abcd', 2) 返回的值是 “ab” 或 “cd” 在字符串中第一次出现的位置。*divisor* 必须是正整数, 必须把 *needle* 分成均匀的长度。*needle* 或子串在 *haystack* 中不存在, 函数值为 0。函数列表中该函数的函数名为 CHAR.INDEX (3)。

(3) CHAR.LENGTH (*strexpr*) 数值型函数。函数值为自变量 *strexpr* 值的以字符为单位并去掉尾部空格后的长度。

(4) CHAR.LPAD (*strexpr*, *length*) 字符型函数。返回一个字符串, 在字符串表达式的左侧增加空格, 扩展到 *length* 所规定的长度。*length* 必须是正整数, 其范围为 1~255。在函数列表中, 此函数名为 CHAR.LPAD(2), 意为两个自变量。

(5) CHAR.LPAD (*strexpr*, *length*, *char*) 字符型函数。与前一个相同, 但不是用空格, 而是用 *char* 变量代表的字符串完整复制在 *strexpr* 代表的字符串左侧扩展。*char* 必须是用单引号括起的字符串常量。此函数在函数列表中名为 CHAR.LPAD(3), 意为 3 个自变量。

(6) CHAR.MBLEN(*strexpr*, *pos*) 数值型函数。返回字符表达式 *strexpr* 代表的字符在 *pos* 指定位置上的字符所占的字节数。

(7) CHAR.RINDEX (*haystack*, *needle*) 数值型函数。返回一个整数, 它表明字符串 *needle* 在字符串 *haystack* 中最后出现的开始位置。返回 0 表示字符串 *needle* 不在 *haystack* 中。例如 CHAR.RINDEX (*var1*, 'abcd') 返回整个字符串 “abcd” 在自变量 *var1* 的值代表的字符串中最后一次出现的位置。此函数在函数列表中名为 CHAR. RINDEX (2), 意为两个自变量。

(8) CHAR.RINDEX (*haystack*, *needle*, *divisor*) 数值型函数。返回一个整数, 它表明字符串 *needle* 在字符串 *haystack* 中最后出现的开始位置。返回 0 表示字符串 *needle* 不在 *haystack* 中。第 3 个自变量是可选择的, 它是一个整数, 用来表示将字符串 *needle* 平均分成被查询的字符串的数目。它必须是一个可以将字符串 *needle* 整分的正整数。没有第 3 个自变量, 功能与上一个函数相同。此函数在函数列表中名为 CHAR. RINDEX (3), 意为 3 个自变量。

例如, CHAR.RINDEX(*var1*, 'abcd', 1) 最后的参数 1 表明, 把 “abcd” 分成单独的一个个的字符, 函数值为任何一个字符在自变量 *var1* 值代表的字符串中最后一次出现的位置。CHAR.RINDEX(*var1*, 'abcd', 2) 的函数值是 “ab” 或 “cd” 最后出现在 *var1* 中的起始位置。

CHAR.RINDEX(*var1*, 'abcd', 2) 最后的参数 2 表明, 把 “abcd” 分成长度相等的两部分 “ab” 和 “cd”, 函数值是这两个字符串中任何一个在自变量 *var1* 值代表的字符串中最后一次出现的位置。此函数在函数列表中名为 CHAR. RINDEX (3)。

(9) CHAR.RPAD (*strexpr*, *length*) 字符型函数。返回字符串, 它的长度由 *length* 决定: 在字符串表达式的右侧加空格, 以达到 *length* 的长度, *length* 的值必须在 1~255 之间。此函数在函数列表中名为 CHAR. RPAD(2), 是两个自变量的函数。

(10) CHAR.RPAD (*strexpr1*, *length*, *strexpr2*) 三个自变量的字符型函数。返回字符串。第 3 个变量 *char* 是可选的, 没有第 3 个自变量, 函数功能与上一个函数相同。函数值是在字符串的右侧增加若干自变量 *strexpr2* 代表的字符串, 达到自变量 *length* 指定的长度。*strexpr2* 必须是一个带有引号的字符串或其值是字符串的表达式。此函数在函数列表中名为 CHAR. RPAD(3), 意为 3 个自变量。

(11) CHAR.SUBSTR (*strexpr*, *pos*) 字符型函数。函数值为自变量 *strexpr* 代表的字符串中从 *pos* 开始到其结尾处的子字符串。此函数在函数列表中名为 CHAR.SUBSTR(2)。

(12) CHAR.SUBSTR (*strexpr*, *pos*, *length*) 字符型函数。函数值为自变量 *strexpr* 代表的字符串中从 *pos* 开始, 长度为 *length* 的子字符串。此函数在函数列表中名为 CHAR.SUBSTR(3)。

(13) CONCAT (*strexpr*, *strexpr* [...]) 字符型函数。函数中每个自变量都是一个字符串表达式。该函数返回一个字符串, 它是各自变量代表的字符串按括号中的顺序串接起来的结果。此函数要求两个或两个以上字符型自变量。

(14) LENGTH(*strexpr*) 数值型函数。返回 *strexpr* 代表的字符串以字节为单位的长度。对

于 Unicode 码的字符串变量,它是每个自变量值的字节数,不包括尾部空格。但对编码页面模式,它就是定义的,包括尾部空格的变量长度。在编码的页面模式下,要得到以字节为单位的除去尾部空格的长度,需要使用嵌套函数调用 `LENGTH(RTRIM(strexpr))` 来求得。

(15) `LOWER(strexpr)` 字符型函数。函数值为将自变量 *strexpr* 中的大写字母改变为小写字母,其他字符不变。自变量可以是字符串变量、字符串表达式,也可以是字符串常量。例如变量 *name* 的值是 *Jery*, `LOWER(strexpr)` 的值为 *jery*。

(16) `LTRIM(strexpr)` 字符型函数。函数值为自变量 *strexpr* 值去掉首部空格的结果。在函数列表中的函数名为 `LTRIM(1)`。

(17) `LTRIM(strexpr[,char])` 字符型函数。函数值为自变量 *strexpr* 的值去掉首部变量 *char* 值代表的字符。第 2 个自变量 *char* 的值必须是单个字符。在函数列表中的函数名为 `LTRIM(2)`。

(18) `MBLEN.BYTE(strexpr,pos)` 数值型函数。函数值是自变量 *strexpr* 在 *pos* 指定位置以字符为单位的字节数(如英文字符是 1 字节,中文是 2 字节)。

(19) `NORMALIZE(strexpr)` 字符型函数。函数值是自变量 *strexpr* 的标准化版本。在 Unicode 模式中函数值是 Unicode NFC。对页面方式无效,函数值就是自变量值,但长度可能与输入的长度不同(Unicode 国际统一编码标准)。

(20) `NTRIM(varname)` 函数值是自变量 *varname* 没有去掉尾部空格的值,自变量 *varname* 的值必须是一个变量名,不能是表达式。

(21) `REPLACE(a1,a2,a3)` 字符型函数。在 *a1* 代表的字符串中所有 *a2* 字符串都用 *a3* 字符串代替。自变量 *a1*、*a2*、*a3* 必须在函数调用前处理成字符串值。例如, `REPLACE("abcabc","a","x")` 函数值为 *"xbcxbc"*。在函数列表中该函数名为 `REPLACE(3)`。

(22) `REPLACE(a1,a2,a3[,a4])` 字符型函数。在 *a1* 代表的字符串中的 *a2* 字符串用 *a3* 字符串代替 *a4* 次。可选的自变量 *a4* 指定替换发生的次数。自变量 *a1*、*a2*、*a3* 必须在函数调用前处理成字符串值(字符串变量或者括在引号中的字符串常量)。可选的自变量 *a4* 必须处理成非负整数。例如 `REPLACE("abcabc","a","x",1)` 函数值为 *"xbcab"*。在函数列表中该函数名为 `REPLACE(4)`。

(23) `RTRIM(strexpr)` 字符型函数。返回截取了尾部空格后的字符串。该函数在函数列表中名为 `RTRIM(1)`。

(24) `RTRIM(strexpr,char)` 字符型函数。函数值是自变量 *strexpr* 的值截取了尾部 *char* 代表的字符后的字符串。*char* 必须是一个带有引号的单个字符或其值是单个字符的字符表达式。该函数在函数列表中名为 `RTRIM(2)`。

(25) `STRUNC(strexpr,length)` 字符型函数。函数值是自变量 *strexpr* 截取 *length* 指定的长度(字节为单位),然后去掉尾部空格。

(26) `UPCAS(strexpr)` 字符型函数。函数值为字符串表达式 *strexpr* 值中小写字母变为大写后的字符串。

## 16. 时间间隔生成函数(Time Duration Creation)

时间间隔生成函数共 4 个,见第 5 章。

## 17. 时间间隔提取函数(Time Duration Extraction)

时间间隔提取函数共 8 个,见第 5 章。

## 1.6 获得帮助

### 1.6.1 SPSS 帮助系统

单击各窗口的【帮助】按钮就可以打开系统【帮助】菜单,如图 1-30 所示,可获得多项帮助。不同窗口的【帮助】菜单内容略有不同,主要的帮助内容如下。

(1) 单击【帮助→主题】打开【帮助】对话框,如图 1-31 所示。在【目录】窗口的【搜索】栏内输入关键字并单击【执行】按钮,即可按关键字搜索;或者单击【搜索范围】打开【建立搜索范围】对话框,建立自己的搜索范围等。操作类似于 Windows 系列软件的帮助系统,不再赘述。



图 1-30 系统【帮助】菜单

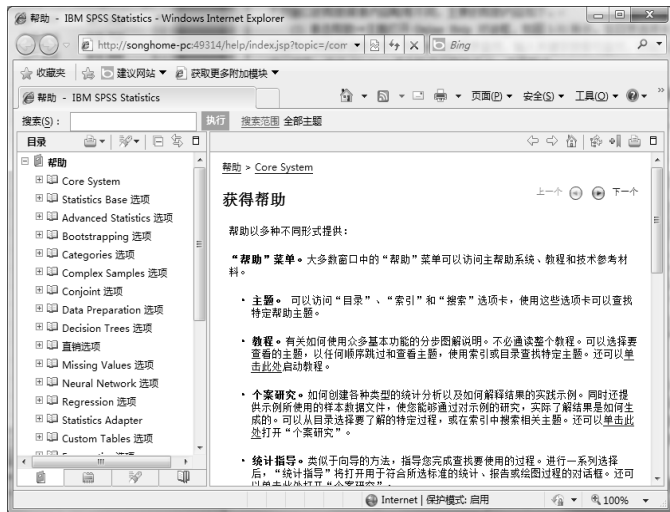


图 1-31 【帮助】对话框

(2) 单击【帮助→教程】打开教程帮助系统,对初学者而言是入门向导,见图 1-32。在图 1-32(a)中,单击十字图标,可以一层层打开树形目录。单击一个菜单项,可以获得如图 1-32(b)的指导画面。指导窗口分两部分,左面是用图解释,右面是文字说明。单击右下角的左右箭头按钮◀▶可以向前、向后翻页;可以按树形目录查找需要的帮助信息;单击指导窗口右上角的【房子】按钮,回到帮助主页。可以在主页目录中改换其他帮助方式。

(3) 单击【帮助→个案研究】选项进入另一个英文的【个案研究帮助】对话框,它是对各种分析过程的操作指导,内容很丰富,有例题,有操作步骤,有选项说明,还有输出结果的解释及结论的得出。操作方法与教程相同。这个菜单只有在数据窗口和输出窗口的【帮助】菜单中才有,见图 1-33。

(4) 单击【帮助→统计辅导】打开统计学指导系统,是对基本统计方法的指导。

指导内容按照下面的思路 and 结构解决读者在初步学习统计分析时可能出现的问题。即要分析什么、数据是什么类型、要什么样的输出、操作步骤以及对话框的详细说明。

在窗口中,左面窗口保持树形目录,便于查找;右面窗口根据需要指导的内容分两列列出读者可能存在的问题。

①【您希望做什么?】左列列出要进行分析的内容,供读者选择,右面一列对应左面的项目为显示示例输出,一级窗口要求选择要进行的分析,列出了指导的主要内容,见图 1-34(a)。单击目录窗口中树形结构中的一项,或者在右窗口左列中选择一个主题,打开如图 1-34(b)所

示的帮助窗口。可以单击一项标题，然后单击【下一个】按钮看其解题过程，单击【显示示例输出】，可以看到例题输出。对选择的每一项分析，都可以单击“下一个”旁边的向右箭头按钮，进入下面各级对话窗口。

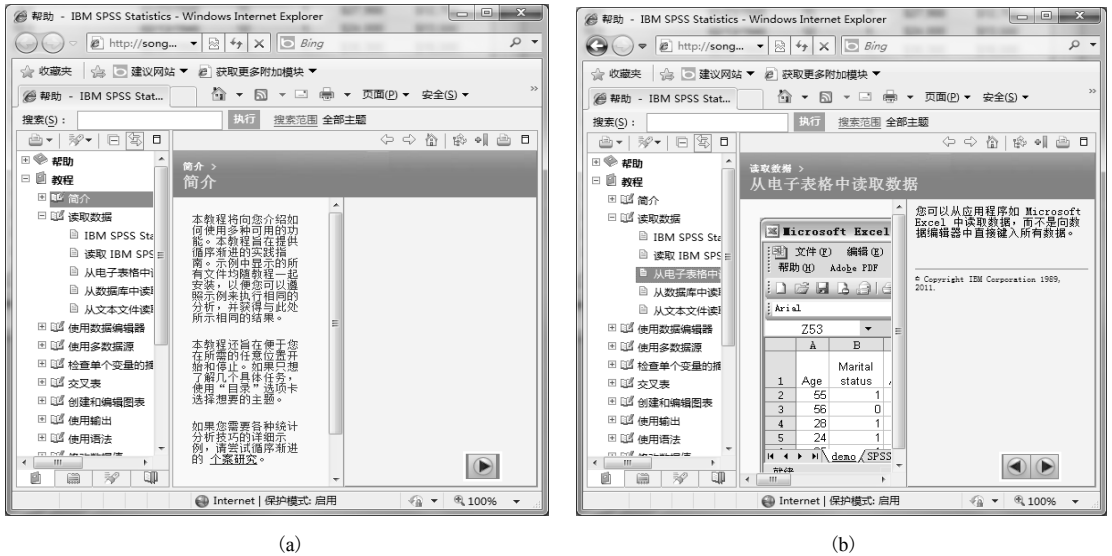


图 1-32 教程的菜单和指导内容示例

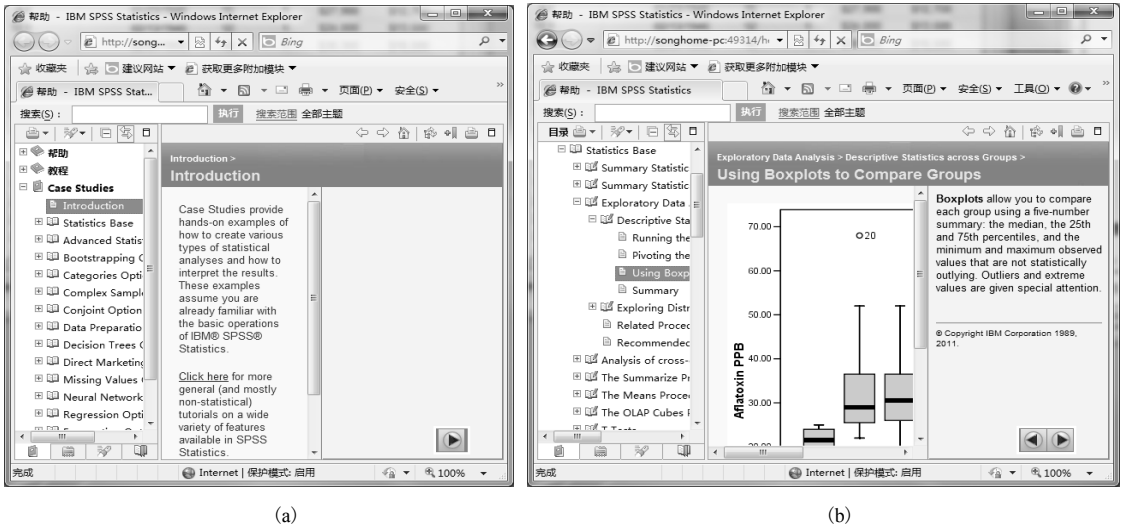
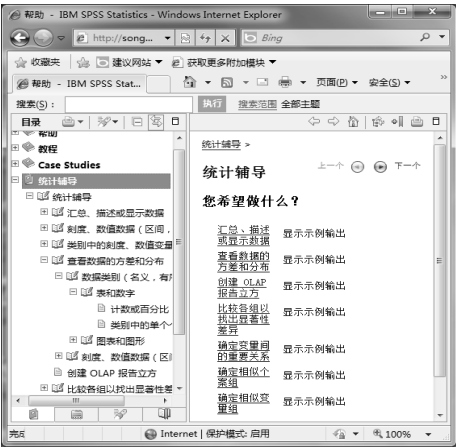


图 1-33 个案研究菜单和指导内容

②【您要汇总哪种类型的数据?】二级窗口中，列出可以分析的数据类型，帮助功能对各种类型数据可以进行进一步的阐述；在左侧显示各种可能的数据类型。选择一种数据类型，单击【下一个】按钮，显示对数据的说明；单击【显示示例输出】，针对指定的数据类型，给出例题输出。

③【您需要哪种显示?】三级窗口中，列出可能的输出，是表和数字，还是图表和图形，帮助对各项输出做详细的说明，单击【显示示例输出】，给出例题输出。

(5) 关于过程语句的帮助系统。SPSS 窗口操作方式使操作变得容易，但是包含的方法和选项有限。需要使用语句补充分析功能和窗口运行方式没有包括的分析过程。对高级分析方法，语句的帮助信息就显得更重要。语句帮助信息显示在 Adobe Reader 阅读器窗口中，见图 1-35。



(a)



(b)

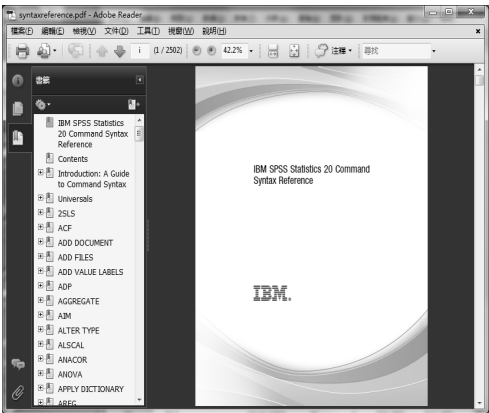
图 1-34 统计辅导的菜单窗口及使用统计学指导的帮助窗口

说明:

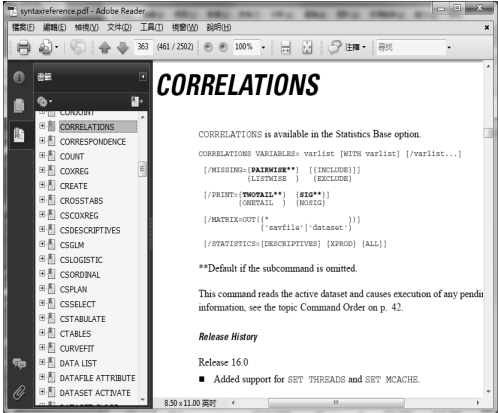
- (1) 因为语句的帮助文件是 PDF 格式的文件, 阅读语句的帮助信息需要安装 Adobe Reader。
- (2) 语句的帮助文件全部用英文书写。

操作方法是单击【帮助→指令语法参考】自动打开 Adobe Reader, 语句帮助信息显示在 Acrobat Reader 阅读器窗口中, 见图 1-35(a)。

窗口左边是语句帮助信息的菜单, 单击一个具体的过程语句名, 右窗口显示具体的语句帮助信息, 见图 1-35(b)。



(a)



(b)

图 1-35 语句帮助窗口

## 1.6.2 右键帮助

### 1. 对话框中的右键帮助

在对话框的变量表中, 用右键单击一个变量, 出现小菜单, 见图 1-36(a), 包括以下 3 组 6 项。在小菜单中组间用横线隔开。

- (1) 【显示变量名称】。
- (2) 【显示变量标签】。变量表显示变量名还是变量标签, 在系统参数设置中已经设置好, 这里可以改变。

- (3) 【按字母顺序排列】。
- (4) 【按文件顺序排序】。即按变量在数据文件中出现的顺序排列。
- (5) 【按度量水平排列】。即按测度水平排列。改变变量的排列顺序，便于查找。
- (6) 【变量信息】。选择这一项，打开【变量信息】对话框，给出变量详细信息，包括【值标签】的下拉列表，见图 1-36(b)。这些帮助信息有助于选择分析变量。



(a)

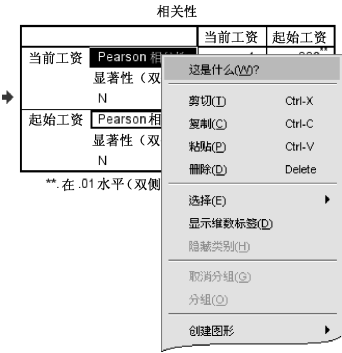


(b)

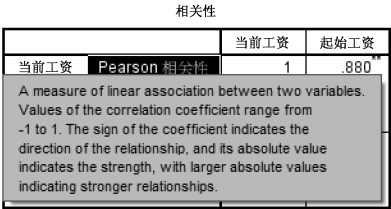
图 1-36 对话框中变量的右键帮助

2. 输出表格中的右键帮助

在输出窗口中双击一个表格，激活它。在表格输出的某个统计量上，单击右键都会出现一个列表，见图 1-37(a)，选择第一项【这是什么?】，会给出对该项统计量的解释，见图 1-37(b)。



(a)



(b)

图 1-37 输出表格的右键帮助

习 题 1

1. IBM SPSS Statistics 软件有几种运行方式？什么是混合运行方式，它有什么特点？
2. IBM SPSS Statistics 有几种类型的窗口，每个窗口的主要功能是什么？
3. 什么是输出窗口(或语句窗口)的主窗口，什么是主窗口的标志？怎样把非主窗口变成主窗口？分别叙述主窗口和非主窗口的作用，以输出窗口为例说明之。
4. 通过什么菜单项设置系统参数？
5. IBM SPSS Statistics 的统计分析功能分布在何处？
6. 从何处可以获得帮助信息？系统提供的帮助有几种形式？

# 第2章 数据与数据文件

## 2.1 变量定义与数据编辑

### 2.1.1 数据编辑器

SPSS 启动后，屏幕上出现的是 SPSS 的数据编辑器，也称数据窗口，如图 2-1 所示。读者在该窗口可建立、打开数据文件。为便于建立变量和查看变量属性，数据窗口分为数据视图、变量视图两种，其组成与功能如下。

(1) 窗口标题栏。当 SPSS 启动后，屏幕显示窗口名称为“未标题 1[数据集 0]-IBM SPSS Statistics 数据编辑器”。随着打开查看的【增加标题】栏中的【未标题】后面的数字增长，数据集后的数字也在增长。当窗口中的数据已经保存到数据文件时，窗口标题栏则显示该窗口中的数据文件名。窗口标题栏下面是菜单栏和工具栏。



图 2-1 数据编辑器的两个窗口

(2) 【变量视图】窗口用于定义和编辑变量的属性，见图 2-1(a)。变量显示区是一个二维表格。左面是行号，也是变量的序号。变量的属性显示在平面表格的第一行，包括变量的变量名、类型、宽度、小数位数、变量标签、值标签、读者自定义缺失值、显示格式(对齐方式、显示列数)和测度方式及变量在分析中的角色等。

(3) 【数据视图】窗口用于输入、显示和编辑数据，见图 2-1(b)。

在工具栏下面是数据栏与数据输入栏：左边一栏是当前数据栏，显示当前光标位置上的变量名和当前记录号。右边一栏为数据输入栏，显示光标位置上的数据值。从键盘输入的数据先显示在此栏中，回车后系统根据定义的变量长度选择合适的形式显示在光标定位的单元格中。数值过大或过小，都有可能使用科学计数法显示数据。

数据显示区：是一个二维平面表格，左面的行号即观测序号；在表格顶部显示变量名，在



它下面的各单元格中显示各变量值。被选定的单元格边框色加深，单元格有底色。所选定单元格中的数据值显示在数据输入栏中。

2.1.2 定义变量

输入数据之前首先要定义变量。定义变量即定义变量名、变量类型、变量长度(小数位数)、变量标签(或值标签)和变量的格式(显示宽度、对齐方式、缺失值标记等)、缺失值和测度方式。

定义变量的步骤如下。

1. 进入定义变量状态

单击【变量视图】选项卡，使数据编辑窗口置于定义变量状态，如图 2-2 所示，每行定义一个变量。

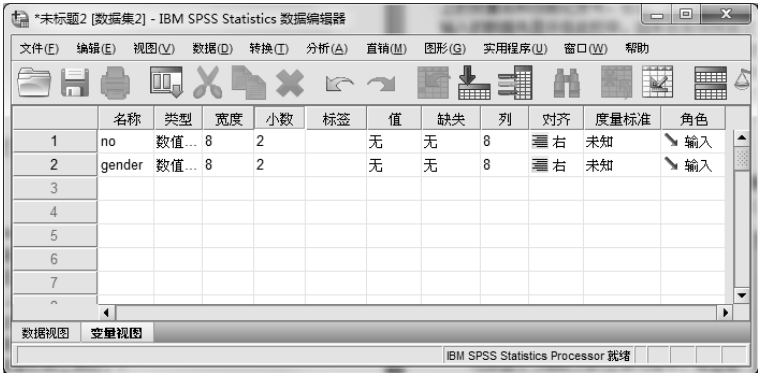


图 2-2 定义变量的窗口

2. 定义变量名

光标置于【名称】列的空单元格中，单击单元格后输入变量名。例如，输入 gender 作为变量名。回车后在同行各单元格中系统自动给出了变量的默认属性。

3. 变量的默认属性值

- (1) 类型。指变量类型，默认类型为数值型。
- (2) 宽度。指变量长度，默认长度为 8。
- (3) 小数。指小数位数，默认小数位数为 2。
- (4) 标签。指变量标签；值，指值标签；缺失，指缺失值定义，这些默认为无，由读者自定。
- (5) 列。指列宽，变量在数据视图中所占列宽默认为 8 个英文字符。
- (6) 对齐。指对齐方式，默认右对齐。
- (7) 度量标准。指测度方式，默认为等间隔测度方式。

如果认为默认的属性与要定义的变量属性不符，可以在同行各属性单元格中设置读者所需要的变量属性。

4. 定义变量类型与宽度

(1) 定义变量类型

单击【类型】列的单元格，默认的数字旁出现删节号。单击删节号，打开【变量类型】对话框，如图 2-3 所示。

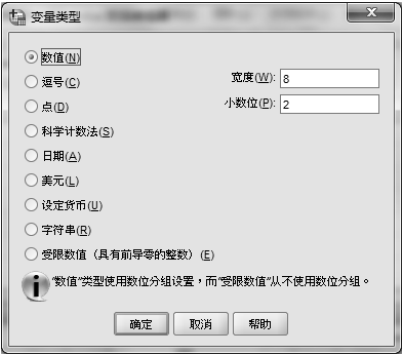


图 2-3 【变量类型】对话框

小数位数，要改变其值，可在单元格中双击鼠标左键，在编辑状态下输入读者认为合适的值；或者用鼠标单击单元格中出现的上下箭头按钮，增加或减少变量宽度值。

5. 定义变量标签

定义变量标签是为了注释变量名含义，可以在输出表格和图中使用，以便理解。在【变量视图】窗口中，双击【标签】相应的单元格，输入注释即可，要尽量简单明了。例如，对 *gender* 变量，可以给出汉字“性别”作为变量的标签。SPSS 16 以上版本软件都可以输入中文标签，每个分析过程的主对话框的原变量表中会在显示英文变量名的同时显示中文标签，使操作变得容易。可以使用【编辑】菜单中的【选项】菜单项的功能设置在输出表格中是否使用在此定义的中文标签，详见 1.3.6 节中的内容。

6. 定义与修改值标签

(1) 定义值标签。单击【值】栏相应的单元格，再单击单元格右侧出现的删节号，打开【值标签】对话框，见图 2-4。在【值】框中输入变量值，在【标签】框中输入对该值含义解释的标签。单击【添加】按钮，一个值标签就被加入到第三个框，即值标签清单中。例如，在定义 *gender* 变量的过程中，数值 1 表示男性，数值 2 表示女性，则先在【值】框中输入“1”，在【标签】框中输入“男”，单击【添加】按钮，列表框中便增加了一个值标签，显示 1=“男”。用同样方法定义第二个值标签，清单中显示 2=“女”，值标签定义完毕。单击【确定】按钮，确认定义的变量标签和值标签正确无误，并返回【变量视图】窗口。定义中文值标签并在【编辑】菜单的【选项】功能(见 1.3.6 节中的解释)中定义在输出表格中使用这个值标签，会使解释输出结果变得更加容易。

(2) 修改值标签。要修改变量的值标签，在【值标签】对话框中，按如下步骤进行。首先在值标签列表中选择要加以修改或删除的值标签表达式，鼠标单击使其反向显示。此时，变量值和该值的标签分别显示在列表上方的【值】、【标签】框中。

删除操作：单击加亮的【删除】按钮，被选定的值标签就从值标签列表中删除。

修改操作：在【值】框中可以输入新的变量值，在【标签】框中输入新标签。例如，选择列表中的 *gender* 变量值 2 的值标签表达式，并在【值】框中修改 *gender* 变量值，将 2 改为 0，标签“女”不变，单击【更改】按钮，列表中的表达式由【2=“女”】改为【0=“女”】，修改完成。

一个值不能定义两个不同的标签；不同的值不能赋予相同的标签。如果用英文定义标签，还可以单击对话框右上角的【拼写】按钮，查拼写错误。

## 7. 读者自定义缺失值

在【变量视图】窗口中，单击变量与“缺失”列对应的单元格，然后单击右侧的【删节号】按钮，打开【缺失值】对话框，见图 2-5，给出读者定义的变量缺失值。



图 2-4 【值标签】对话框

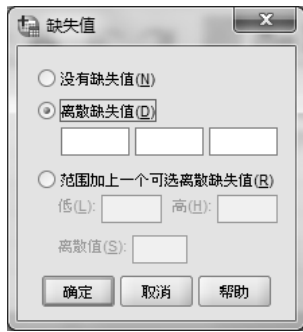


图 2-5 【缺失值】对话框

先选择一种缺失值的类型，再进行具体定义。定义用户缺失值的类型有 3 种：

(1) 【没有缺失值】。本选项是系统的默认状态。如果当前变量的值测试、记录完全正确，没有遗漏，则可选择此项。

(2) 【离散缺失值】。选择这种方式，可以在下面的 3 个矩形框中输入 3 个可能出现在相应的变量中的缺失值，也可以少于 3 个。在进行统计分析时系统遇到这几个值，则作为缺失值处理。例如，对于性别变量，如果定义了用 1 表示男，用 2 表示女，则值 0、3、4 都被认为是非法的。可以将这 3 个值分别输入到 3 个矩形框中，当数据文件中出现这几个数据时，系统将按缺失值处理。

(3) 【范围加上一个可选离散缺失值】。选择此项后，除了【低】和【高】参数框外，还有【离散值】，即范围以外的一个值。例如，如果定义变量 *height* 的值中输入的错误数据有 1.40、1.90、1.95 和 2.03，而且在 1.90~2.03 之间没有正确的身高测试值，正确值在大于 1.40 和小于 1.90 的范围内，则可选择此种定义缺失值的方式。在【低】参数框中输入 1.90，在【高】参数框中输入 2.03，在【离散值】框中输入 1.40。

如果这 3 种定义缺失值的方式都不能把所有的非法值包括在内，则要在数据文件中查出错误数据并进行修改，修改成系统缺失值。或者在【语句】窗口中利用缺失值函数解决定义缺失值的问题。

## 8. 定义变量的显示格式

(1) 定义显示用的列宽度。在【变量视图】窗口中，单击【列】相应的单元格，再单击出现的上下箭头按钮。增加或减少【列】宽度值，如图 2-6(a) 所示。

(2) 定义显示时的对齐方式。在【变量视图】窗口中，与变量行对齐相应的单元格中显示的是默认的对齐方式。对数值型变量，系统默认右对齐；对字符型变量，系统默认左对齐。如果要改变默认的对齐方式，则单击对齐列相应的单元格，有 3 种可选择的方式：【左对齐】、【居中对齐】、【右对齐】，在下拉列表中任选一种，见图 2-6(b)。

(3) 关于默认值。

① 字符串(字母数字)变量默认类型为名义变量。

② 带有【值标签】的数值型变量默认类型为有序变量，即等级变量或称定序变量。

- ③ 没有定义【值标签】的数值型变量，但数值的个数少于指定的数量则被设置成序号。
- ④ 没有定义【值标签】的数值型变量，但数值的个数大于选项功能中指定的数，则被设置成度量型，即等间隔测度的变量或称尺度变量。


9. 定义变量测度类型

在【变量视图】窗口中，与变量行度量标准列相应的单元格中显示的是默认的变量测度方式度量，即等间隔测度。

默认变量值的数量为 24。若要改变这个值，则选择【编辑→选项】菜单项，在对话框中的数据【选项】卡的指定测量级别栏中设置这个值。


如果要改变默认的测度类型，则单击【度量】列相应的单元格，打开下拉列表，如图 2-6(c)所示。在下拉列表中有 3 个可选择的类型。

(1) 【度量】。尺度变量，对等间隔测度的变量或者表示比值的变量选择此项，如身高、体重。




(a)

(2) 【序号】。定序变量，对其值表示顺序的变量选择此项，如比赛名次、职务、职称等，可以是数值型变量，也可以是字符型变量。




(b)

(3) 【名义】。标称变量，它是分类变量的一种，可以是数值型变量，也可以是字符型变量。例如，变量值是对所喜欢颜色的回答，表示宗教信仰、党派等的变量。



(c)

(4) 【角色】。定序变量，对其值表示顺序的变量选择此项，如比赛名次、职务、职称等，可以是数值型变量，也可以是字符型变量。



(d)

图 2-6 定义变量的列格式和测度方式

10. 定义变量的角色

有些对话框使用预先在数据编辑器中定义的变量角色。例如在【自动线性建模】的【字段】选项卡对话框中就有“使用自定义角色”这样的选项。在打开这样的对话框时，满足角色要求的变量会自动显示在目标列表中。默认的变量角色是输入变量。角色定义只影响支持角色分配的对话框，对语句命令语法没有影响。

要改变变量的角色只须单击【角色】列相应的单元格，展开如图 2-6(d)所示的下拉列表。选择下拉列表中的某项角色定义即可。可以选择的角色有六种：

- 【输入】角色。预设变量在分析中作为输入变量（例如回归分析中的自变量、预测变量）。
- 【目标】角色。预设变量在分析中将作为输出变量或目标变量，例如回归分析中的因变量。
- 【两者】。预设变量在分析中可以同时作为输入变量和输出变量。
- 【无】。不预设变量在分析中的角色。
- 【分区】角色。预设变量用于将数据划分为单独的训练样本、检验和验证样本。
- 【拆分】角色。预设变量作为拆分文件的变量。定义了拆分角色，在【数据】→【拆分文件】功能的【分割】对话框中会自动显示在【分组方式】变量表中。

11. 确认全部定义的属性

经过上述操作，定义完一个变量的属性参数。可以重复上述操作，定义其他变量属性参数。所有变量名及其属性都显示在【变量视图】窗口中。如果对定义的属性满意，则按数据视图选项卡，转移到数据编辑窗口，输入数据。

2.1.3 定义日期变量

【定义日期】功能可产生周期性的时间序列日期变量，还可以给时间序列分析的输出加标签。按【数据→定义日期】顺序打开【定义日期】对话框，如图 2-7 所示。在该对话框中选择各项与建立、修改、删除日期型变量有关的操作。

1. 关于【个案为】栏

【个案为】栏即日期类型选项栏，其中各项都是定义日期变量的时间间隔和为定义时间变量做准备的功能项。利用该对话框建立具有一定时间间隔的日期变量必须满足下列条件：

- (1) 在数据窗口中已经有一个数据文件。
- (2) 在该数据文件中的变量名不能与将要建立的日期变量的默认变量名重名，否则新建日期变量将覆盖同名变量。系统默认变量名有：YEAR\_、QUARTER\_、MONTH\_、WEEK\_、DAY\_、HOUR\_、MINUTE\_、SECOND\_和 DATE\_。

(3) 对于每个【个案为】列表中所列的功能项，SPSS 生成若干数值型变量，新变量名以下画线结尾。同时生成一个字符型变量 Date\_，用以解释生成的日期变量。例如，如果在【个案为】列表中选择了“星期、日、小时”，则生成 4 个新变量 WEEK\_、DAY\_、HOUR\_和 DATE\_。



图 2-7 【定义日期】对话框

2. 【当前日期】栏

该项在【个案为】列表下方，显示项与定义新日期变量有关。在当前日期标题下面显示的是已经存在的与即将生成的日期变量同名的变量及其定义，以提醒注意。

3. 【第一个个案为】栏

定义起始日期值，该值作为第一个观测，接下来的各观测值根据时间间隔自动生成。

4. 【更高级别的周期】栏

显示项与【个案为】栏中所选择的项目对应，在【更高级别的周期】栏指定相应的重复周期。例如一年中的月数，一周中的天数。在可以输入数值的区域后面显示的是可以输入的最大值。

图 2-7 所示的是在【个案为】栏中选择了星期、日、小时，则在【第一个个案为】栏中显示了将产生的 3 个日期型变量的值，这些值在【数据编辑器】中作为第一个观测。更高级别的周期显示的是各变量的最大值，即该对应变量的周期。

5. 【个案为】栏中的主要功能

- (1) 【未注日期】(倒数第 2 项)。选择此项将删除当前数据文件中与系统默认的日期型变量名相同的变量，为使用【个案为】栏中的某些功能项的执行创造条件。
- (2) 【设定】(倒数第 1 项)。该功能指出由命令语句生成的日期变量，而非使用【定义日期】功能生成的日期变量。如每周 4 个工作日的日期变量，它只反映当前工作的数据文件状态，对数据文件没有影响。

除以上两项功能外全部都是生成日期变量的功能项。

【例 1】 生成日期变量。

以定义年、季度、月为例说明操作方法。

假定已经在【数据编辑器】中建立了一个变量 no,【标签】为编号,且输入了 1~20 的值,数据集中有了 20 个观测。

- (1) 按【数据→定义日期】顺序单击菜单项,打开【定义日期】对话框。
- (2) 在【个案为】栏内选择【年、季度、月份】项。在【第一个个案为】栏内显示:

①【年】框显示 1900,这是系统默认数值,改变该值输入 2008,见图 2-7,此项表明第一观测的 YEAR\_变量值为 2008。

②【季度】框显示变量的起始值为 1,周期为 4,按 1、2、3、4 顺序排列;输入 3。

③【月】框显示变量的起始值为 1,周期为 12。输入 8。

单击【确定】按钮,在数据窗口中生成的新变量有: YEAR\_、QUARTER\_、MONTH\_和对这 3 个变量值的解释变量 DATE\_,如图 2-8 所示。



(a) (b)  
图 2-8 日期变量生成的结果(数据窗口与变量窗口)

第一个观测, no 值为 1, YEAR\_、QUARTER\_、MONTH\_三个变量的值分别为 2008、3、8, DATE\_的值为 AUG2008。下一个观测, 4 个变量的值分别为 2008、3、9、SEP2008。共生成 20 个观测。


从图中可以看出,要想使用【定义日期】功能自动生成日期变量,则在原数据文件中的各观测必须都是按某时间顺序取得的,而且时间顺序必须与【定义日期】对话框中【个案为】栏目中的某一选项相对应才行。

2.1.4 数据录入与编辑

1. 录入数据

输入数据的操作方法是多种多样的,可以定义一个变量后便输入这个变量的值(纵向进行),也可以定义完所有变量后,按观测来输入(横向进行)。

【数据编辑器】的二维表格中顶部标有变量名,左侧标有观测序号。一个变量名和一个观测序号就指定了唯一的单元格。可以使用上下左右箭头将插入点光标(当前单元格的定位)移动到相邻的位置;用 Home、End 键将插入点光标移动到同行首单元格或同行尾单元格。也可以使用滚动条或 PgUp、PgDn 上下移动一屏。

单击【值标签】图标按钮。所有设置了值标签的变量均显示值标签，图 2-9 所示是显示值标签的状态下录入变量 *gender* 数据。当输入一个变量值时在单元格中单击向下的箭头，在下拉列表选择一个定义过的值标签，这样只使用鼠标即可输入有值标签的变量值。

2. 编辑数据

如果知道某个变量的某个值输入错误，只要定位到相应的单元格，重新输入这个数据即可。

(1) 移动指针到指定序号的观测

当数据量很大时，要在【数据视图】中查找一个观测是很麻烦的事。可以利用【转向个案】功能解决此问题。例如，要修改第 108 观测的性别，先把鼠标指针移到数据视图中【性别】变量的任意一个观测上。按【编辑→转向个案】顺序单击鼠标，或单击工具栏上的图标按钮，如图 2-10(a)所示，打开【转到】对话框的【个案】选项卡，如图 2-10(b)所示。在【转向个案数】栏中输入要查找的观测号，例如输入 108。单击【转向】按钮。第 108 行的某个变量(性别)值被加深显示，可以即刻修改。不关闭对话框，还可以继续查找。如果输入的观测号大于数据文件观测个数，则指针停留在最后一个观测上。

(2) 查找变量

按【编辑→转向变量】顺序单击鼠标，或单击工具栏上的【转向变量图标】按钮，如图 2-11(a)所示。打开【转向】(SPSS 汉化为【转到】)对话框的【变量】选项卡，单击【转向变量】栏的向下箭头，在下拉菜单中选择要查找的变量名，例如选择【职务分类】变量，见图 2-11(b)。单击【转向】按钮，【职务分类】变量列所有值被加深显示，不关闭对话框还可以继续查找。

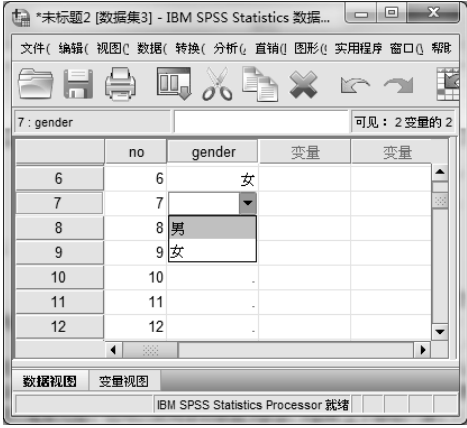


图 2-9 显示值标签的变量

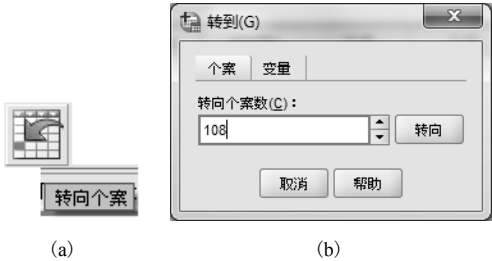


图 2-10 查找观测对话框及查找结果

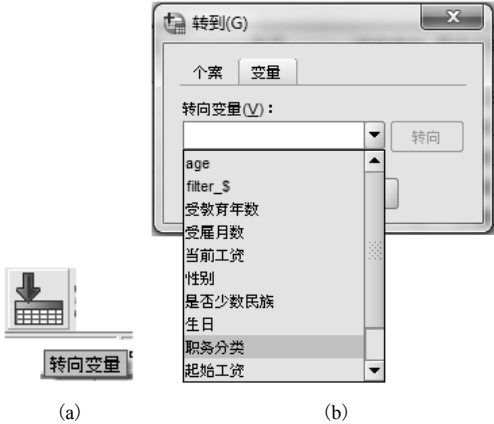


图 2-11 查找变量对话框及查找结果

【转到】对话框有两个选项卡——【个案】和【变量】，可以根据查找目标进行切换。

(3) 在【数据视图】窗口中查找或替换指定变量中的指定数据(定位到单元格)

【例 2】 查找 *educ* 值为 19 的观测。

- ① 鼠标指针移至变量 *educ* 受教育年所在列中的任意单元格，单击鼠标，指定在该列中查找。

② 按【编辑→查找】顺序单击鼠标,或单击工具栏上的【查找】图标按钮,见图 2-12(a)。打开【查找和替换】对话框,该对话框标题栏显示要查找值的所属变量【列:受教育年数】,见图 2-12(b)。

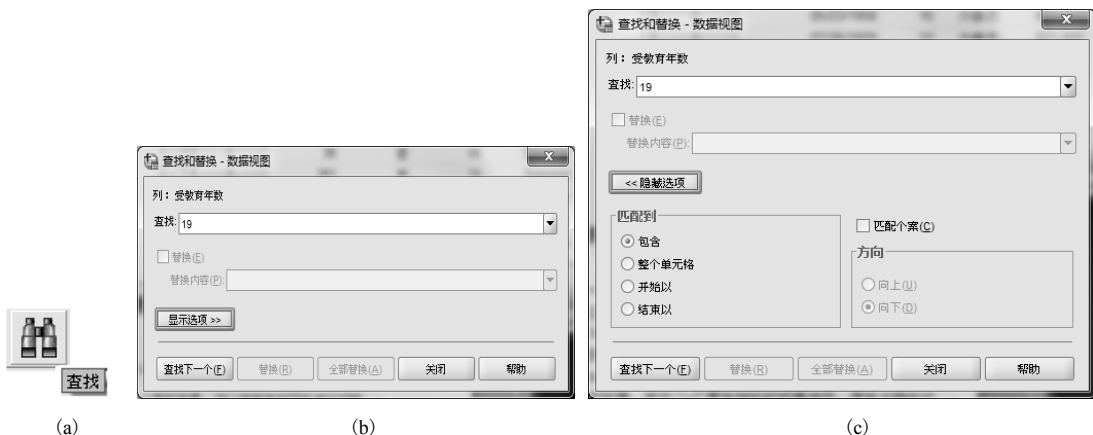


图 2-12 查找数据对话框及查找结果

③ 在【查找】框中输入要查找的变量数值。本例要求查找【受教育年数=19】的观测,输入 19。

④ 单击【查找下一个】按钮,即向观测序号大的方向查找;找到后加深显示查找内容,见图 2-12(b)。再单击【查找下一个】按钮,可以继续查找。直到显示提示信息,查找终止。单击【关闭】按钮,退出对话框。

#### (4) 匹配查找

单击【显示选项】按钮打开另一半窗口,见图 2-12(c)。在【匹配到】栏中选择一种方式,对查找目标进一步定义:

- 【包含】。查找包含指定内容的变量值。例如,在【姓名】变量中找姓张的,查找栏填入“张”,选择此项,所有名字中有“张”的会一个个显示出来。
- 【整个单元格】。必须整个单元格的内容完全与指定内容一致才算找到。
- 【开始以】。查找以指定内容开头的变量值。例如,选择此项查找\$123,找到的内容可以是\$123.0、\$1234.0,但是不包括\$1,23.0。
- 【结束以】。查找以指定内容为结尾的变量值。

注意:日期时间变量在【数据视图】窗口中按显示格式查找。例如,显示格式为 10/18/2008,查找 10-18-2008 就找不到。在数据窗口中只能向观测号大的方向查找,不能向观测号小的方向查找。

#### (5) 替换功能

当查找到一个观测符合查找标准时,替换选项加亮。选择【替换】项,在【替换内容】后面的矩形框中填写替换的内容。当查找到一个目标时,单击【替换】按钮,则查找到的内容被替换。再单击【查找下一个】按钮,找到一个再单击【替换】按钮,再替换一个。若单击【全部替换】按钮,则所有与查找内容匹配的都被替换成【替换内容】框中填写的内容。

#### (6) 在【变量视图】窗口中的查找与替换

在【变量视图】窗口中,只能对变量名、变量标签、值、缺失值和自定义变量属性列进行查找。且只能对标签、值的内容和自定义的变量属性列进行替换。

注意:在【值】列可以对值和值标签进行查找,但是要替换数据值就会把原来的值标签一起删除了。




### (7) 插入一个变量

如果要在现存变量的右边界左面增加一个变量，只需单击【变量视图】选项卡标签，转换到【变量视图】窗口，在变量表最下面一行定义新变量。


如果想把要定义的变量放在已经存在的变量之间，可进行如下操作：

① 确定插入位置。在【数据视图】窗口中将指针置于要插入新变量的列中任意单元格上，单击鼠标左键；或者在【变量视图】窗口中，单击新变量要占据的那一行的任意位置。

② 单击【编辑→插入变量】命令，或单击【插入变量】图标按钮，在选定的位置上插入一个变量名为“Var0000n”的变量，其中“n”是系统给的变量序号。原来占据此位置的变量及其后的变量依次后移。

③ 切换到【变量视图】窗口中，对插入的变量定义属性，包括更改变量名。然后切换到【数据视图】窗口，输入该变量的数据。

### (8) 插入一个观测

观测的排列无关紧要，其排列次序可以用排序功能整理。如果确实需要插入一个观测，可以将指针置于要插入观测的一行的任意单元格中，单击鼠标。单击【编辑→插入个案】命令，或单击工具栏上【插入观测】图标按钮实现。结果在选中的一行上增加一个空行，可以在此行上输入该观测的各变量值。

### (9) 变量和观测的删除、复制和移动




在【数据视图】窗口中单击变量名，或者在【变量视图】窗口中单击变量所在的行号就选择了一个变量；对变量的删除和移动可以在这两个窗口中进行。因为不允许有同名变量，所以变量不能复制。

对观测的删除、复制和移动只能在【数据视图】窗口中进行。单击一个行号就选择了这一行上的观测。



移动变量(或观测)只需在选择要移动的对象后，单击【编辑】菜单中的【剪切】命令，找到插入位置，先插入一个空变量(或空观测)，单击空变量的变量名(或空观测序号)，即选择这个空变量(或空观测)，然后单击【编辑】菜单中的【粘贴】命令，就将剪贴板中的变量(或观测)粘贴到空变量(或空观测)的位置上了。

要复制观测，只需把上述步骤中的剪切改为单击【编辑】菜单的【复制】命令即可。

要删除变量或观测，只需选择要删除的对象后，按 Delete 键或者单击【编辑】菜单中的【清除】命令。

另外，工具栏中图标按钮、、的作用即剪切、复制、删除，使用方法与上述命令一致。

### (10) 恢复删除或修改前的数据

单击【编辑】菜单中的【撤销】命令，或单击工具栏中的【撤销】图标按钮可撤销前一步操作。单击【编辑】菜单中的【重新】命令或单击图标按钮，可恢复撤销前的状态。

## 2.1.5 根据已有的变量建立新变量

### 1. 使用计算变量功能完成对新变量值的计算

在进行数据的分析处理时，往往需要根据已经存在的变量建立新变量。这一工作可以直接通过 SPSS 语句实现。对 SPSS 来说，体现其特点的更直观方法是通过【计算变量】对话框完成。

(1) 按【转换→计算变量】顺序，打开如图 2-13(a)所示的【计算变量】对话框。



图 2-13 【计算变量】对话框和【计算变量：类型和标签】对话框

(2) 在【目标变量】框中输入目标变量的名称，用来接收计算的值。目标变量名可以是一个新的变量名或一个定义过的变量名。如果是新变量，则单击【类型与标签】按钮，打开【计算变量：类型和标签】对话框，可定义新变量的类型和标签如图 2-13 (b) 所示。

- ① 【标签】栏，为新变量指定标签。
    - 【标签】。可在该框中输入长达 120 个字符的说明变量含义的标签。
    - 【将表达式用作标签】。利用表达式的前 110 个字符作为标签。
  - ② 【类型】栏。为变量指定类型。只有两种基本类型可以指定：【数值】型，这是默认设置。【字符串】型，要在【宽度】参数框中输入字符串的宽度。
- 单击【继续】按钮返回【计算变量】对话框。

(3) 在【数字表达式】框中组合合理的数学表达式。对话框的软键盘中包含了常数、数学运算符、关系表达符号、逻辑运算符。【数字表达式】矩形框相当于计算器的显示屏。在【数字表达式】框中可以利用鼠标或键盘进行相应的编辑操作，方法如下：

- ① 在左面的矩形框中选择已经存在的变量，移入【数字表达式】框中。
- ② 在操作板上选择数字或运算符，单击后出现在【数字表达式】框中。
- ③ 在【函数组】框中选择需要的函数，双击选中的函数；或单击选中的函数，然后单击向上箭头按钮，使选中的函数出现在【数字表达式】框中。函数自变量用问号表示。
- ④ 移动插入点“I”形光标至函数名称后面的括号中，然后按①所示的方法选择自变量并单击【向右箭头】按钮，使其置于括号之中，代替括号中表示自变量的问号。

- (4) 表达式组成规则参见 1.5.2 节的内容，另外需要注意：
- ① 自变量必须放在函数名后的括号中。
  - ② 每一个关系表达式必须单独完成，例如，把年龄变量分段： $age1=3$  (if  $age \geq 30 \ \& \ age < 40$ ) 与  $age1=4$  (if  $age \geq 40 \ \& \ age < 50$ )，定义变量  $age1$  的两个值，两个值分别以变量  $age$  的不同值为条件确定，则必须分两步完成。
  - ③ 圆点“.”是表达式中唯一合法的小数点符号。

(5) 条件表达式(If)。

当不同特点的观测使用不同的表达式计算新变量的值时，新变量的值需要分步进行计算。在【计算变量】对话框中确定计算部分新变量值的表达式后，再利用条件表达式选择观测。对使条件表达式值为真的观测，使用【计算变量】对话框中确定的表达式计算新变量的值；对那些使条件表达式为假或缺失的观测，新变量的值或缺失值，或保持不变。

① 在【计算变量】对话框中单击【如果】按钮，打开【计算变量：If 个案】对话框，如图 2-14 所示。



图 2-14 【计算变量：If 个案】对话框

② 根据需要选择下列选项：

- 【包括所有个案】。包括数据集中所有观测，这是默认选项。选择此项对所有观测使用【计算变量】主对话框中的计算表达式来计算新变量的值，没有任何条件。
- 【如果个案满足条件则包括】。只对满足条件表达式的观测才计算新变量的值。选择此项后，激活其下面的矩形框，输入条件表达式。操作方法与【计算变量】对话框中的操作方法相同。

③ 条件表达式规则：

大多数条件表达式至少要包括一个关系运算符，并且可以通过关系运算符来连接多个条件表达式。例如：

`age >= 21` 表示只有 `age` 大于等于 21 的观测才会被选择。

`Salary*3 < 100000` 表示只有 `Salary` 乘以 3 的值小于 100000 的观测才会被选择。

`Salary*3 < 100000 & jobcat <> 3` 表示只有 `Salary` 乘以 3 小于 100000 并且 `jobcat` 不等于 3 的观测才会被选择。


逻辑运算符连接的两个关系表达式必须单独完成，例如 `age >= 18 & age < 35` 合法，而 `age >= 18 & < 35` 非法。

④ 单击【继续】按钮表示确认输入的条件表达式并返回主对话框。

(6) 单击【确定】按钮,对符合【计算变量: If 个案】对话框中所设置条件的观测,按主对话框中确定的计算表达式计算新变量的值。

## 2.1.6 打开、保存与查看数据文件

### 1. 打开一个已有的数据文件

按【文件→打开】顺序单击鼠标,或单击工具栏上的图标按钮,打开【打开数据】对话框。在【查找范围】框中指定文件存储位置。数据文件类型栏显示为 SPSS Statistics (\*.sav)。找到或输入要打开的数据文件名,双击之,就可以将数据文件显示在数据窗口中。

在【打开数据】对话框中,单击【文件类型】框内的向下箭头,打开 SPSS 允许打开的文件类型列表。数据文件的类型大致有以下几种。

- SPSS (\*.sav): SPSS 建立的数据文件,扩展名为“\*.sav”。
- SPSS/PC+ (\*.sys): SPSS/PC 或 SPSS/PC plus 建立的语句文件,扩展名为“\*.sys”。
- Systat (\*.syd, \*.sys): SYSTAT 建立的数据文件(扩展名为“\*.syd”)或语句文件(扩展名为“\*.sys”)。
- 便携 (\*.por): 用 SPSS 简便格式保存的数据文件。
- Excel (\*.xls): Excel 建立的表格数据文件。SPSS 可以直接打开 Excel 电子表格文件。
- Lotus (\*.w\*): 用 Lotus 1-2-3 格式写的数据库文件。可以是 1A 版、2 版、3 版 Lotus1-2-3 记录的数据文件。它的一行转换成一个观测,变量是一列。
- Sylk (\*.slk): 用 Sylk 格式保存的数据文件。
- dBASE (\*.dbf): 数据库格式文件,扩展名为“\*.dbf”。可以是各种版本 dBASE 或 FoxBase 建立的数据库文件。一个记录转换成数据窗口中的一个观测。
- SAS (\*.sas7bdat, \*.sd7, \*.sd2, \*.ssd01, \*.ssd04, \*.xpt): 各版本的 SAS 软件生成的数据文件。
- Stata (\*.dta): 各种版本的 Stata 软件生成的数据文件。
- 文本格式文件 (\*.txt, \*.dat, \*.csv)

### 2. 保存数据文件

保存数据文件可以使用【文件】菜单中的【保存】和【另存为】命令。操作方法与 Windows 系列应用软件的文件保存方法一样,而 SPSS 数据文件可以选择不同变量保存为不同的文件。

可选择的数据文件的类型很多。基本上可以打开的文件类型都是可以保存的文件类型,但大部分会丢失变量标签和值标签。保存为文本文件时有下面两种类型:

- 以制表符分隔 (\*.dat)。保存为 ASCII 码文件,用制表符作为两个观测之间的分隔符。  
如果一个软件不能读取其他任何格式的数据文件,可以使用此种格式保存数据。在将数据保存为此种格式文件的同时,变量标签、值标签、缺失值定义均丢失。
- 固定 ASCII 格式 (\*.dat)。保存为固定列格式的 ASCII 码文件。

注意: \*.dat 文件不是标准格式文件,在 SPSS 中是文本格式的数据文件。在 Windows 操作系统中被默认为视频文件。因此要注意打开和保存时的操作。

### 3. 保存部分变量

单击【文件→另存为】,打开如图 2-15 所示的【将数据保存为】对话框,在【查找范围】栏中设置保存位置,在【文件名】栏中输入文件名。

单击【保存】对话框中的【变量】按钮，打开如图 2-16 所示的【数据保存为：变量】对话框。在该对话框中选择要保存的变量。系统默认全部保留，所有变量名前都标有对钩。只有标有对钩的变量才被保存到文件中。选择不要保存的变量，即去掉该变量前面的对钩，激活【全部保留】按钮。单击【全部保留】按钮，所有变量都被选中，则【全部保留】按钮变暗。或者选择【全部丢弃】按钮，去掉所有变量前面的对钩，可以重新选择要保存的变量。单击【继续】按钮，返回【将数据保存为】主对话框，单击【保存】按钮。完成保存部分变量的操作。



图 2-15 【将数据保存为】对话框

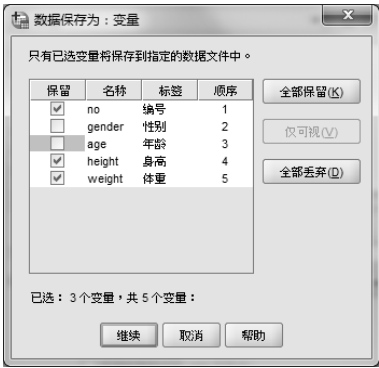


图 2-16 【数据保存为：变量】对话框

【例 3】 另存为 ASCII 码数据文件实例。

数据文件 data02-02.sav 中有 5 个变量：*no*(编号)、*gender*(性别)、*age*(年龄)、*height*(身高)、*weight*(体重)。把数据保存为固定格式的 ASCII 码文件的操作如下：


- (1) 按【文件→另存为】顺序单击鼠标左键，打开【将数据保存为】对话框。
- (2) 在【将数据保存为】对话框的【查找范围】栏中指定存储位置(驱动器、目录)，在【保存类型】栏中选择文件类型为固定 ASCII 格式\*.dat，即其扩展名为\*.dat，并输入文件名保存。在输出窗口中显示保存记录，如表 2-1 所示。

表 2-1 文件以 ASCII 码形式存入指定位置

Variable	Rec	Start	End	Format
<i>no</i>	1	1	2	F2.0
<i>gender</i>	1	3	4	F2.1
<i>age</i>	1	5	7	F3.2
<i>height</i>	1	8	12	F5.2
<i>weight</i>	1	13	15	F3.2

表 2-1 中第 1 列是变量名。第 2 列是该变量所在的记录号，这些变量处于同一个记录中。第 3 列是对应的变量所占的起始列号，第 4 列是对应的变量所占的结束列号，变量 *no* 占两列，变量 *gender* 占两列……第 5 列是对应的变量的格式。前两个变量是字符型，后 3 个变量是数值型。由于各变量值间没有空格，如果在其他软件中打开此文件，则该表对重新整理数据很重要，应该保存。

4. 查看变量信息

如果数据集变量很多，在【数据编辑器】中查看变量全面信息就比较困难，操作烦琐。简便办法是，在数据窗口中选择一个变量，单击【实用程序】菜单中的【变量】命令，打开【变量】对话框，如图 2-17 所示(数据文件 data02-02.sav)；也可以单击【变量】图标按钮，打开该对话框。

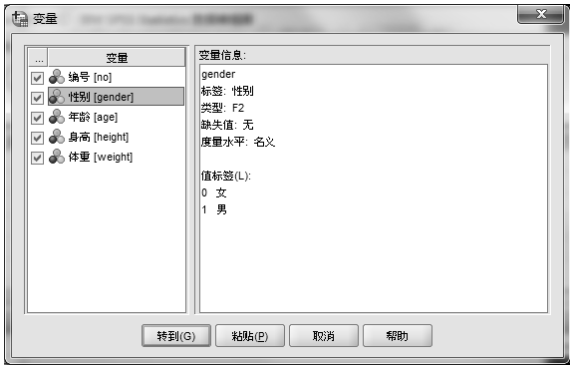


图 2-17 【变量】对话框

对话框中左半部是【变量】列表，列出当前【数据编辑器】中定义的所有变量名和测度类型图标。鼠标单击【变量】列表中的一个变量，右半部分【变量信息】显示区列出指定变量的属性。【变量信息】框中只能显示一个变量的属性信息。例如，图 2-17 所示是【性别】变量的信息，第一行是变量名，gender；第二行是变量标签：性别；第三行是变量类型，类型：F2，表示 2 位的数值型变量；第四行是变量的缺失值定义，缺失值：无，说明如果该变量值为 0，则认为是缺失值。空行后是值标签：值 0 标签为女；值 1 标签为男。

单击【转到】按钮，关闭【变量】对话框，返回【数据编辑器】窗口。

5. 查看文件信息

可以利用【文件】菜单的命令查看所有定义的变量。方法是按【文件→显示数据文件信息】顺序单击菜单项，在二级菜单中：

(1) 单击【工作文件】。当前数据窗口中所有变量的有关信息均显示在输出窗口变量信息表中。

(2) 单击【外部文件】，打开【显示外部数据集信息】对话框，指定一个外部数据文件，文件中所有变量信息均显示在输出窗口中。

文件信息包括文件保存位置、文件类型、生成日期，以及是否定义了加权变量等。变量信息包括变量在数据编辑窗口中的位置序号、变量名、变量标签、值标签、格式和缺失值。

该功能对查找一个已经建立的数据集，看是否是想要打开的数据集很有用。经过上述操作，只要到输出窗口查看即可。

2.2 数据文件的转换

2.2.1 ASCII 码数据文件的转换

几乎所有有计算功能或管理数据功能的软件，都可以输出 ASCII 码数据文件，因此掌握 ASCII 码数据文件转换成 SPSS 数据文件的方法是非常重要的。

1. 不同格式的 ASCII 码数据文件

SPSS 可以读入 ASCII 码数据文件，并将其转换为 SPSS 格式，显示在数据窗口中。ASCII 码数据文件有固定宽度格式和使用分隔符的自由格式两种。所谓固定格式，即一个观测(或称

记录)占一行或若干行，每个变量所占起始列和结束列是固定的，如图 2-18 所示。自由格式即每个变量在文件中的列位置不一定是固定的，各变量值之间使用相同的符号(如空格或逗号)隔开，转换时根据分隔符和变量值排列顺序进行。

(1) 在固定格式排列的 ASCII 码数据文件中，数据的排列方式有以下两种。

① 每行安排一个观测，每个变量值之间由空格分隔，见图 2-18(a)。这种固定格式也可以看作使用分隔符的自由格式数据文件，见数据文件 data02-02a.txt。

② 每行安排一个观测，但变量值之间没有任何分隔，如图 2-18(b)所示，数据安排实例见数据文件 data02-02b.txt。

(2) 使用分隔符的自由格式 ASCII 码数据文件。

① 每行安排若干个观测，或整齐地排列两个观测，如图 2-19(a)所示。数据文件 data02-03.txt 看上去既是固定格式 ASCII 码数据文件，又是使用分隔符的自由格式 ASCII 码数据文件。转换程序将其归为使用分隔符的自由格式文件。如果使用固定格式转换操作，那么在转换后通过对数据文件的编辑才能形成正确的转换结果。

② 每行一个或多个观测，使用分隔符将各变量值分开，甚至一个观测从一行中间开始并在下一行继续，如图 2-19(b)所示，见数据文件 data02-04.txt。

2. 固定格式 ASCII 码数据文件的转换

以图 2-18(b)为例，说明固定格式 ASCII 码数据文件转换为 SPSS 数据文件的操作。

图中所示数据文件为 data02-02b.txt，数据由 5 个变量组成，变量编号占第 1、2 列，性别占第 3 列，年龄占第 4、5 列，身高占第 6~9 列(小数点占一列)，体重占第 10、11 列。



图 2-18 不同排列的固定格式 ASCII 码数据文件

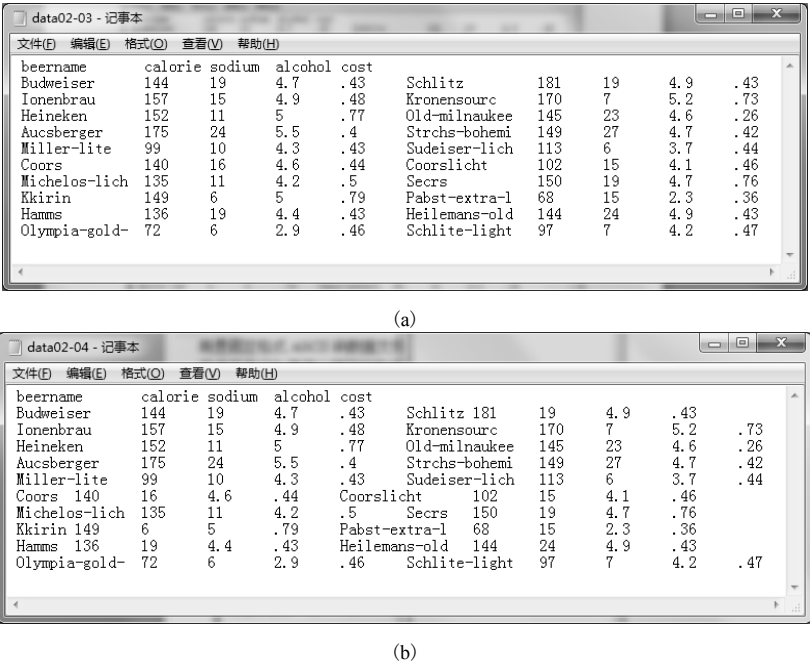


图 2-19 不同排列的自由格式 ASCII 码数据文件

数据文件转换步骤如下：

(1) 按【文件→打开文本数据】顺序打开【打开数据】对话框，在【文件类型】下拉菜单中选择【文本格式(\*.txt、\*.dat、\*.csv)】项，矩形框中将列出这一类数据文件。指定一个扩展名为 txt 的数据文件(dada02-02b.txt)并单击【打开】按钮，打开【文本导入向导-第 1 步，共 6 步】对话框，如图 2-20 所示。分 6 步完成转换工作，此为第 1 步。数据显示在下面带有标尺的预览框内。



图 2-20 【文本导入向导-第 1 步，共 6 步】对话框

择【固定宽度】选项。每个选项后面都有解释，仔细阅读并确定你的文本文件是哪种排列。

②【变量名称是否包括在文件的顶部?】

【是】或【否】选其一。移动滚动条可以看到，顶部没有变量名，因此选择【否】。单击【下一步】按钮打开【文本导入向导-第 3 步(共 6 步)固定宽度】对话框，如图 2-22 所示。

(3) 在【文本导入向导-第 3 步(共 6 步)固定宽度】对话框中要求提供有关观测的信息，一个观测相当于数据库中的一个记录。各选项的含义与操作如下：

①【第一个数据个案从哪个行号开始?】参数框。要求指定数据文件中第一个包括数据值的行号，默认值为 1。如果在顶行包括了对变量的解释文字或变量标签，则该值不能是 1。如果没有变量行，但是只需分析一部分数据，也可能不从第一个观测开始，则应该在该项后面的数值栏中设置具体值。

右面的栏询问【您的文本文件与预定义的格式匹配吗?】，如果需要，则选中【是】，并通过单击【浏览】按钮指定一个扩展名为 .tpf 的文件。通常不选择该选项。单击【下一步】按钮，打开【文本导入向导-第 2 步(共 6 步)】对话框，如图 2-21 所示。

(2) 在【文本导入向导-第 2 步(共 6 步)】对话框中回答两个问题：

①【变量是如何排列的?】

选择【分隔】项，即是使用分隔符将变量隔开的；选择【固定宽度】，则意为是使用固定列宽的。图 2-21【预览】框中的数据排列整齐，列宽是固定的，因此选



图 2-21 【文本导入向导-第 2 步(共 6 步)】对话框



②【多少行表示一个个案?】参数框。要求回答一个观测占几行,以确定何处为一个观测的结束位置和下一个观测的起始位置。本例每个观测占一行。

③【您要导入多少个个案?】栏,指定要转换的观测数。

- 【全部个案】。指定转换所有观测。此为默认的选择。
- 【前 $n$ 个个案】。指定前  $n$  个观测,  $n$  是自定义的正整数。框中输入  $n$  值。
- 【个案的百分比】。指定一个百分比,转换系统按指定的百分比随机提取观测。由于随机采样是通过对每个观测产生一个独立的伪随机数进行的,因此该百分比是一个近似值,最后采样得到的样本占观测总数的百分比接近这个指定值。本例指定转换所有观测,选择第一项。



图 2-22 【文本导入向导-第 3 步(共 6 步)固定宽度】对话框

(4)【文本导入向导-第 4 步(共 6 步)固定宽度】对话框如图 2-23 所示。在浏览窗口中,

加竖线将各变量值分开,标明将如何读取数据。浏览区上有尺,左有观测号。插入分隔线和去除分隔线的方法有二:

① 在需要插入变量分隔线处单击鼠标左键,出现分隔线。本例的 ASCII 码数据文件中各变量值间没有分隔符,因为 1、2 列为编号值,第 3 列为性别值,因此要在第 2、3 列之间加分隔线。在列间单击鼠标左键,就会插入一根分隔线;右键单击一根已经存在的分隔线,则该分隔线变成蓝色,再单击【删除终止】按钮,则删除该分隔线。

图 2-23 【文本导入向导-第 4 步(共 6 步)固定宽度】对话框

② 列号。若在第  $n$  列右侧需要加分隔线,则输入该列的列号  $n$ 。单击【插入终止】按钮,则第  $n$  列右侧加入分隔线,单击【删除终止】按钮,将第  $n$  列右侧分隔线删除。

由计算机产生的连续数据流,各变量值之间没有空格或其他分隔符,很难确定一个观测从哪里开始,到哪里结束,应该使用其他应用程序将其重新编辑成便于转换的排列。

本例在第 2、3、5、9 列号上加了分隔线,见图 2-23。

(5)【文本导入向导-第 5 步(共 6 步)固定宽度】对话框如图 2-24 所示。这一步确定变量名和变量类型。对话框【数据预览】栏内显示根据第 4 步的变量分隔线划分的各变量数据。变量



图 2-24 【文本导入向导-第 5 步(共 6 步)】对话框

使用吗?】栏内选择【是】,指定将该格式保存到一个文件中,以便对相同或类似数据文件进行转换时使用。单击【另存为】按钮,打开相应的对话框,指定保存位置和文件名;否则,选择【否】。

② 在【您要粘贴该语法吗?】栏内选择【是】,即把各步确定的转换参数粘贴到语句窗口形成命令文件,以便进行类似转换工作时使用;否则,选择【否】。

一切参数设置工作完成后,单击【完成】按钮,转换开始,见图 2-25。最后在数据编辑窗口中显示转换结果,见图 2-26,图中所示为数据编辑器的变量视图窗口和数据视图窗口。在数据编辑窗口中对各变量的标签、值标签、缺失值等属性再进行完善和修改。

名为系统默认的  $V_n$ 。 $n$  为自左至右的变量顺序号。转换程序据此对各变量进行读取并转换成 SPSS 数据文件。

① 在【预览】栏中单击要定义的默认变量名。在变量名称下面输入自己命名的变量名。除应该符合变量名的有关规定外,不能重名。

② 在【数据格式】下拉列表中选择一种类型,定义选中变量的数据类型。

图 2-24 中已经定义了第一个变量  $no$  和第 2 个变量  $gender$ ,第 3 个变量的变量名已经输入完成。

(6) 【文本导入向导-第 6 步(共 6 步)】对话框如图 2-25 所示。

① 在【您要保存此文件格式以备以后



图 2-25 【文本导入向导-第 6 步(共 6 步)】对话框

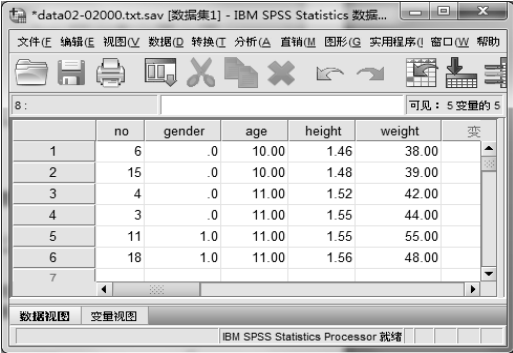


图 2-26 在数据编辑的两个窗口中的转换结果

3. 自由格式 ASCII 码数据的转换

自由格式 ASCII 码数据的文件有以下特性：

- ① 各观测中的各变量值按相同顺序排列，但同一变量的值不一定占有相同的列位置。
- ② 两个值之间以空格、逗号或其他符号分隔。
- ③ 每行可以有不止一个记录(一个记录即一个观测)。

在转换时，读完最后一个定义的变量的值就读完了数据文件中的一个观测，然后 SPSS 读下一个值时就认为是下一个观测的第一个变量的值。因此，定义的变量数必须与 ASCII 码数据文件中的变量数目相同，否则转换后的结果是混乱的。当存在两种类型的变量时，会出现数据与变量类型不匹配的错误。

**【例 4】** 数据文件 data02-04.txt 是一组关于 12 盎司啤酒中的成分和价格的 ASCII 码数据文件，见图 2-19(b)，是一个空格分隔、一行两个记录的文本文件，包括 beername(啤酒名)、calorie(热量卡路里)、sodium(钠含量)、alcohol(酒精含量)、cost(价格)共 5 个变量；空格做分隔符，且空格数不定。从文本文件看，顶部包括变量名。

转换为 SPSS 格式数据文件的操作步骤如下：

(1) 按【文件→打开文本数据】顺序打开【打开数据】对话框，指定一个扩展名为 txt 的数据文件 data02-04.txt，并单击【打开】按钮，打开【文本导入向导】对话框，如图 2-27 所示。分 6 步完成转换工作，此为第 1 步。数据显示在预览框内。可以看出，数据间空格做分隔符，每行一个记录，但排列较乱。

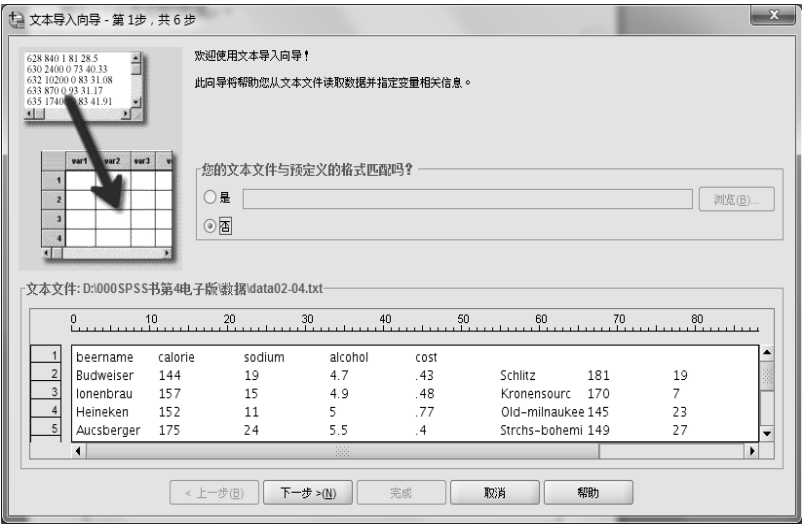


图 2-27 【文本导入向导-第 1 步，共 6 步】对话框

右面的【您的文本文件与预定义的格式匹配吗?】回答该问题，可以选择【是】或【否】。如果选【是】，应该通过单击【浏览】按钮指定一个扩展名为 tpf 的文件。通常不选择该选项。单击【下一步】按钮，打开【文本导入向导-第 2 步(共 6 步)】对话框，如图 2-28 所示。

(2) 在【文本导入向导-第 2 步(共 6 步)】对话框中回答两个问题：

- ① 【变量是如何排列的?】有两个选项：【分隔】，用分隔符隔开变量值；【固定宽度】，固定(列)宽度。

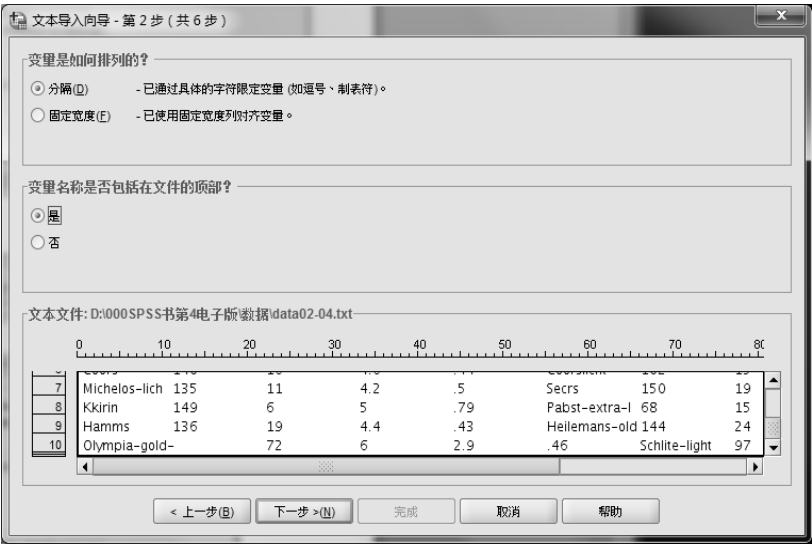


图 2-28 【文本导入向导-第 2 步(共 6 步)】对话框

图 2-27 预览框中的数据排列凌乱，不是固定列宽度，但每两个数据之间均有空格，是使用分隔符的，因此选择【分隔】选项。



图 2-29 【文本导入向导-第 3 步(共 6 步)分隔】对话框

有变量名，所以数据从第 2 行开始。参数框内数值应为 2。

②【如何表示个案?】栏实际上是问一个观测占几行? 以便确定何处为一个观测的结束位置和下一个观测的起始位置，本例为两个观测占一行，此项不能确切回答数据文件排列的实际情况，其中有两个选项：

- 【每一行表示一个个案】。即应该观测占一行。即使变量非常多，使得一行非常长，也属于这种情况。如果各行包括的数据量不同，则每个观测包括的变量数由数值最多的行确定。数值较少的观测，多出来的变量值赋予缺失值。
- 【个案的特定编号表示一个个案】。指定每个观测包括的变量数，告诉系统在何处停止

②【变量名称是否包括在文件的顶部?】有两个选项：【是】或【否】选其一。移动滚动条可以看到，顶部没有变量名，从文本文件及图 2-27 数据预览区中都可以看到顶部包括变量名，见图 2-19(b)，因此选择【是】。单击【下一步】按钮打开【文本导入向导-第 3 步(共 6 步)分隔】对话框，如图 2-29 所示。

(3) 在【文本导入向导-第 3 步(共 6 步)分隔】对话框中提供有关观测的信息。

①【第一个数据个案从哪个行号开始?】指定数据文件中第一个包括数据值的行号，默认值为 1。本例顶行

读一个观测，并开始读下一个观测。该选项允许在同一行中有多个观测，或一个观测开始于一行的中部并在下一行继续。系统根据数值个数读取数据，不管行数。因此每个观测必须包括所有变量的数值(缺失值必须使用分隔符指定)，才能正确进行转换。本例一个观测包括 5 个变量，因此选择此项，并设置数值 5。输入变量数以后，【数据预览】窗口中的数据按 5 个变量整理好。

③【您要导入多少个个案?】。在该栏指定要转换的观测数。

- 【全部个案】。指定转换所有观测。此为默认的选择。
- 【前□个个案】。指定前 n 个观测，n 是由读者输入的正整数。
- 【个案的随机百分比(近似值)】。指定一个百分数，转换系统按指定的百分比随机提取观测。由于随机采样是通过每个观测产生一个独立的伪随机数进行的，因此该百分比是一个近似值。本例指定转换所有观测，选择第一项。

单击【下一步】按钮，打开如图 2-30 所示的【文本导入向导-第 4 步(共 6 步)分隔】对话框。



图 2-30 【文本导入向导-第 4 步(共 6 步)分隔】对话框

(4) 【文本导入向导-第 4 步(共 6 步)分隔】指定分隔符和字符串的标识符。

①【变量之间有哪些分隔符?】

栏中列出的分隔符有 5 种：【制表符】、【空格】、【逗号】、【分号】、【其他】。可以同时选择几种，还可以选择【其他】项，并在其后的文本框中输入一个分隔符。根据指定的分隔符，转换后的数据文件状态显示在【数据预览】栏中，可以查看所指定的分隔符是否有误。本例数据中啤酒名变量值中有可能有空格，所以选择【制表符】，【数据预览】窗口中的观察显示结果选择制表符效果最佳。



图 2-31 【文本导入向导-第 5 步(共 6 步)】对话框

【数据预览】栏中，可以查看所指定的分隔符是否有误。本例数据中啤酒名变量值中有可能有空格，所以选择【制表符】，【数据预览】窗口中的观察显示结果选择制表符效果最佳。

②【文本限定符是什么?】即指数据中的字符串值用什么符号表示。下设 4 个选项：【无】(没有限定符)、【单引号】、【双引号】、【其他】，选择【其他】，需要在其后框中输入一个具体的限定标准。本例中，字符串没有加单引号或双引号标识，所以在右栏中选择【无】。

单击【下一步】按钮，打开如图 2-31 所示的对话框。

(5) 【文本导入向导-第 5 步(共 6 步)分隔】的定义每个变量值的变量名和数据格式，以便在进行转换并组成数据文件时读取各变量值。在【数据预览】栏内选择一个变量，对它进行定义。选择时单击要定义的一列数据顶部的原始变量名，原始变量名出现在【变量名称】栏内及其后面。

- ① 【变量名称】。删掉或覆盖原始的变量名，输入自己定义的变量名。
- ② 【数据格式】。在其下拉列表中选择变量类型。

本例定义变量 beername 为【字符型】变量，对字符型变量还要在后面的【字符】栏中输入字符串长度。本例输入最长的字符数 14；本例还定义变量 calorie、sodium、alcohol、cost 为【数值型】。

(6) 最后一步参见图 2-25，只需回答两个问题。

- ① 【您要保存此文件格式以备以后使用吗?】如果需要，则单击【另存为】按钮指定存储位置和文件名。本例选择【否】。
- ② 【您要粘贴该语法吗?】即是否要将其转换为 SPSS 命令语句。本例选择【否】。

选择过后，单击【完成】按钮，系统开始进行转换。转换后的数据出现在数据编辑窗口中，如图 2-32 所示。在数据编辑窗口中对转换后的数据进行编辑，例如调整每个变量所占宽度等。



图 2-32 数据编辑窗口中的转换结果

2.2.2 数据库文件的转换

任何数据库文件，例如 Excel、dBase、FoxBase、FoxPro、Oracle，要使用 SPSS 软件进行分析处理，就必须将数据库文件转换为 SPSS 格式。这里只介绍最常用的 Excel 文件转换成 SPSS 数据文件的快速完全转换的方法。

快速完全转换就是打开对话框选择一种数据库文件直接打开。以打开中国女排档案的 Excel 文件为例：

- (1) 按【文件→打开→数据】顺序打开【打开数据】对话框，建立搜索路径。
- (2) 打开【文件类型】菜单，选择 Excel 类型，选择中国女排档案文件 data02-17.xls。
- (3) 单击【打开】按钮，打开【打开 Excel 数据源】对话框，如图 2-33 所示。在对话框中做如下设置。

- ① 【从第一行数据读取变量名】。对 Excel 文件来说，回答是肯定的，因此选择此项。

②【工作表】。在该项指定工作簿中的工作表和读取数据的范围。默认值是系统对所指定 Excel 文档的分析得出的，工作表是 Sheet1，数据范围是 A1:E15，即默认为第一个工作表的全部数据。

③【范围】。还可以另指定工作表中的数据范围。本例无须再指定。

④【字符串列的最大宽度】。一般数据库文件中的字符串肯定不会大于这个宽度，所以无须选择。

单击【确定】按钮，转换自动进行。结果如图 2-34 所示。

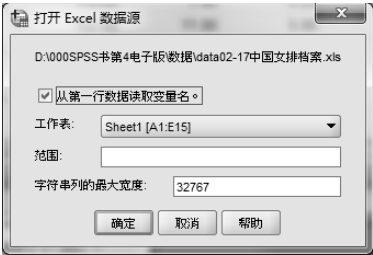


图 2-33 【打开 Excel 数据源】对话框



(a)



(b)

图 2-34 转换结果的变量视图和数据视图

2.2.3 观测的查重

1. 实际工作中有时会输入重复的数据

可能出现的情况如下：

- 同一个观测输入了多次。
- 多个观测共用一个标识变量的值，但是第二标识变量的值不同；例如，同一个家庭的多个成员共用一个家庭地址或家庭编号。
- 标识变量值相同，非标识变量值不同；例如，同一个人或同一个公司，多次或在不同时间购买不同的产品，在记录购买情况的数据文件中，这个人的名字或编号会出现多次。

2. 识别与处理重复观测的方法

(1) 按【数据→标识重复的个案】顺序单击菜单项，打开如图 2-35 所示的对话框。

(2) 定义识别重复观测的根据。

将源变量框中的识别变量移到右边的【定义匹配个案的依据】框中，这些识别变量值相等

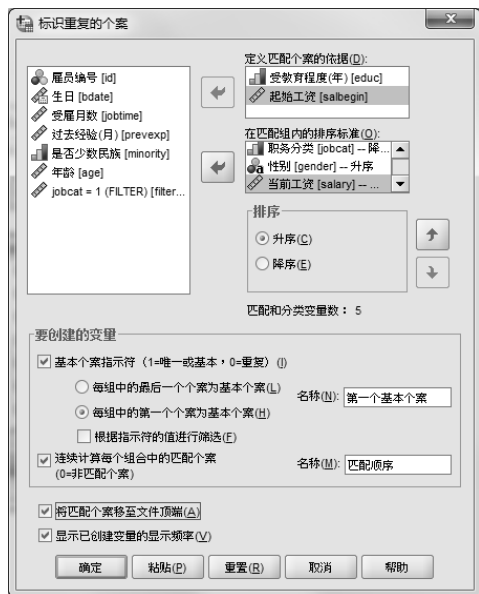


图 2-35 【标识重复的个案】对话框

序变量排序，第一变量值相同的观测形成一组，第一排序变量值相同的组内，按第二排序变量排序……

#### (4) 指定组内排序规则。

在【排序】栏内的两个单项中选择一个：【升序】或【降序】，即决定按排序变量(在匹配组内的排序标准中的变量)值升序还是降序排列组中的观测。

如果指定了两个排序变量，则在【排序】变量栏中选择一个，定义一次【升序】或【降序】。有几个排序变量就定义几次。系统默认按排序变量值的升序排列。

#### (5) 指定指针变量的特性。

对经过查重的数据文件，生成一个指针变量，在【要创建的变量】栏内指定该指针变量如何标识重复观测。

① 【基本个案指示符 (1=唯一或基本, 0=重复)】，选择此项，产生的指针变量的值，对不重复的观测，其值为 1；对重复的观测中的主观测，该变量值也为 1；对重复的观测非主观测(不满足主观测定义的)，其值为 0。

#### ② 定义重复观测组中主观测的条件。

- 【每组中的最后一个个案为基本个案】，选择此项后变量名为最后一个基本个案，在重复观测组中，最后一个观测的指针变量值为 1，其他为 0。
- 【每组中的第一个个案为基本个案】。选择此项后变量名为“第一个基本个案”。在重复观测组中，第一个观测为基本观测的指针变量的值为 1，其他为 0。
- 【根据指示符的值进行筛选】。选择此项后用指针变量作为过滤变量，非主重复观测将从分析中去除，但无须从数据文件中删除。输出的结果和报告与这些观测无关。也就是说，重复的观测只留一个参与后续的分析，根据前两个选择项决定是保留排序后的重复观测的第一个，还是保留最后一个。
- 【连续计算每个组合中的匹配个案】。选择此项，后面的【名称】栏内显示产生的另一个变量匹配顺序，它对有  $n$  个重复观测的组中的各观测标  $1 \sim n$  的值。每个重复观测组

的观测被认为是重复的观测。可以选择一个，也可以选择多个。

如果选择两个以上的识别变量，例如选择了两个识别变量，系统将按第一个识别变量对数据文件排序，第一个变量值相同的，再按第二个识别变量排序。这两个变量值都相同的观测就是读者定义的重复观测。

系统为标识重复观测生成一个变量。对重复观测，该变量的值为 0；对非重复观测，其值为 1。

#### (3) 在匹配组内的排序标准。

系统根据识别重复观测变量找到重复观测后，还要对它们排序。在源变量框内选择一个变量并送入【在匹配组内的排序标准】框内。对符合同一重复条件的观测组按该变量值排序。可以使用排序框右侧的上下箭头改变排序变量的位置。排序变量在该栏中的位置不同，则排序结果页不同。按第一排



自行排列。如果指定了排序变量，则排列顺序取决于排序变量；如果没有指定排序变量，则排列顺序取决于观测在原始数据文件中的顺序。

- **【将匹配个案移至文件顶端】**。选择此项，查重执行的结果会把有重复观测的组移到数据文件的顶部，以便观察。
- **【显示已创建变量的显示频率】**。选择此项，要求生成频数表，包括所生成的新变量各值的计数。例如对主指针变量，频数表给出 0 值的个数和 1 值的个数。1 值的数目表明数据文件中共有多少个无重复的单一观测和主观测。

**【例 5】** 数据文件 data02-01 为有 474 个观测雇员情况的数据。变量：id(雇员编号)、gender(性别)、bdate(出生日期)、educ(受教育年限)、jobcat(职务等级)、salary(当前工资)、salbegin(起始工资)、jobtime(雇用工作月数)、prevexp(以前的工作经历月数)、minority(民族)和 age(年龄)。

使用查重功能查看雇员受教育程度、职务的构成以及初始工资情况。操作步骤如下：

- (1) 按**【数据→标识重复的个案】**顺序单击菜单项，打开对话框。
  - (2) 在源变量框中选择 educ、salbegin 并作为识别变量移到**【定义匹配个案的依据】**框中。
  - (3) 在源变量框中选择 jobcat、gender、salary 并送入**【在匹配组内的排序标准】**框内，作为排序变量。
  - (4) 设置按 jobcat 降序，gender、salary 升序排列。
  - (5) 定义重复观测中开始的一个为主观测，变量名为**【第一个基本个案】**。
  - (6) 要求生成重复观测的顺序变量，选择**【连续计算每个组合中的匹配个案】**项，生成的新变量名为匹配顺序。
  - (7) 选择**【显示已创建变量的显示频率】**。要求生成频数表。
  - (8) 选择**【将匹配个案移至文件顶端】**项。把有重复观测的组移到数据文件顶部。
- 各栏选项安排参见图 2-35。

提交运行后，结果如表 2-2、表 2-3 和图 2-36 所示。

表 2-2 显示，重复的观测共 312 个，主观测 162 个；就是说，162 组中有重复的测量组合。总观测数是 474 个。

表 2-2 指针变量概况表

		频率	百分比	有效百分比	累积百分比
有效	重复个案	312	65.8	65.8	65.8
	主个案	162	34.2	34.2	100.0
	合计	474	100.0	100.0	

表 2-3 是频数分布表。举例说明各项内容：有效值(匹配顺序的值)为 6 的频率是 21 个。

本例指定了两个重复(匹配)依据变量 educ 和 salbegin。一个重复组即第一个重复依据变量值相同(软件中称匹配)，第二个重复依据变量值也相同的观测为一组。该组中的每个观测都给一个序号，即图 2-36 中匹配顺序变量的值。表 2-3 中的有效值就是匹配顺序的值。频率就是匹配顺序值为某有效值的个数。例如，有效值为 6 的有 21 个。这 21 组中重复观测数至少是 6 个，百分比为 4.4%，累计百分比是 72.2%，其含义是，重复的观测数为 6 以上(包括 6)的组共有 21 组，占总观测数的 4.4%，从不重复的观测(有效值为 0)到一组有至少 6 个重复观测的个数占观测总数 474 的 72.2%。

图 2-36 是查重过程运行结束后的数据窗口。为便于查看，删去了与此例无关的变量。新变量 PrimaryFirst 和 MatchSequence 的含义及其各种值的含义也是显而易见的。

表 2-3 重复观测的频数分布表

		频率	百分比	有效百分比	累积百分比
有效	0	94	19.8	19.8	19.8
	1	68	14.3	14.3	34.2
	2	68	14.3	14.3	48.5
	3	42	8.9	8.9	57.4
	4	27	5.7	5.7	63.1
	5	22	4.6	4.6	67.7
	6	21	4.4	4.4	72.2
	7	18	3.8	3.8	75.9
	8	16	3.4	3.4	79.3
	9	14	3.0	3.0	82.3
	10	13	2.7	2.7	85.0
	11	12	2.5	2.5	87.6
	12	10	2.1	2.1	89.7
	13	8	1.7	1.7	91.4
	14	7	1.5	1.5	92.8
	15	6	1.3	1.3	94.1
	16	5	1.1	1.1	95.1
	17	4	.8	.8	96.0
	18	3	.6	.6	96.6
	19	3	.6	.6	97.3
	20	3	.6	.6	97.9
	21	3	.6	.6	98.5
	22	3	.6	.6	99.2
	23	3	.6	.6	99.8
	24	1	.2	.2	100.0
合计		474	100.0	100.0	

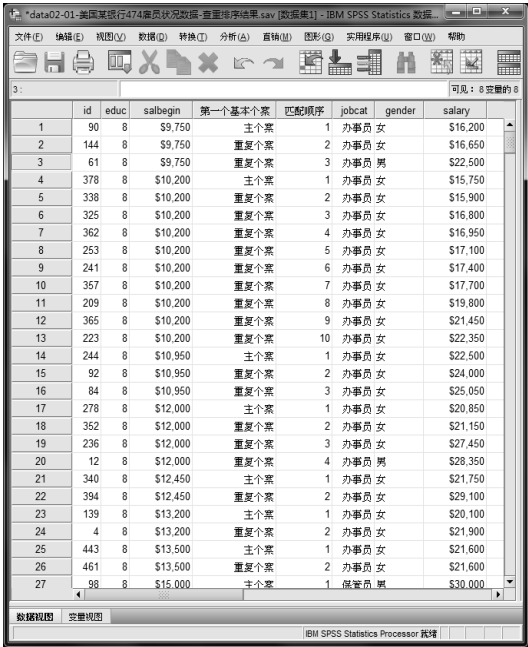


图 2-36 查重后的数据排列

2.3 数据文件操作

2.3.1 数据文件的拆分与合并

1. 数据文件的拆分

在进行数据处理时经常要对数据文件中的观测进行分组分析，但有些分析功能没有设置对分组变量的选项。例如，要使用描述功能(【分析】菜单中)分别求出男生、女生的平均身高。在进行分析之前必须对该数据文件进行拆分。这里的“拆分”并非将一个数据文件拆分为两个或若干个独立的数据文件，而是在同一个数据文件中按某个条件分组。若对数据文件进行了拆分处理，则拆分处理一直有效，直到取消拆分处理或更改拆分变量后，才会有新的变化。关闭 SPSS，也会使拆分失效。具体操作步骤如下：

- (1) 读取数据文件 data02-05。
- (2) 按【数据→拆分文件】顺序打开【分割文件】对话框，如图 2-37 所示。
- (3) 根据对数据的具体需要选择以下选项。
  - 【分析所有个案，不创建组】。这是系统的默认选项。
  - 【比较组】。将各分组的观测数据分别分析，所得的结果放在一起进行比较。
  - 【按组组织输出】。即分别显示各组所得的统计结果。

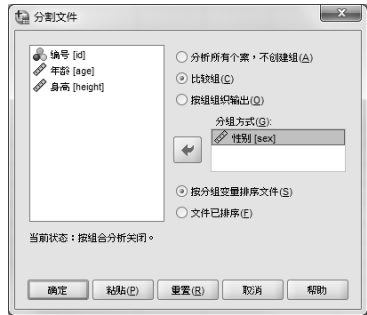


图 2-37 【分割文件】对话框

(4) 从左侧的源变量框中将一个或若干个进行分组的依据变量名选入【分组方式】框中。此处最多可以选择 8 个变量作为拆分变量。这些变量所起的作用相当于排序的 By 变量。

如果只选择了一个变量，以后的分析将会依据该变量的每一个值分为一组，分别进行分析。例如，选择性别变量 sex，分析时分别按 sex=0 和 sex=1 把观测分为两组进行分析。

如果选择了若干个变量，以后的分析将会依据所选择的变量各值的组合分组，对每个组分别进行分析。例如，选择了变量 sex，它有 2 个水平：sex=0，sex=1；还选择了变量 age，它有 3 个水平：age=11，age=12，age=13。分析时分为 6 组进行：sex=0，age=11；sex=0，age=12；sex=0，age=13；sex=1，age=11；sex=1，age=12；sex=1，age=13。

(5) 指明数据文件的当前状态。

①【按分组变量排序文件】。表示要求按所选择的变量对数据文件进行排序，作为拆分文件在分析时才起作用。而从数据窗口中看上去，经过拆分的数据文件与经过同样的变量排序的文件是相同的。如果在进行拆分之前进行了排序，则会节省拆分所需的时间。

②【文件已排序】。表示数据文件已经按所选择的变量排序。

(6) 单击【确定】按钮执行并完成拆分。拆分结果如图 2-38 所示。图 2-38(a)是按 sex 变量值拆分的结果；图 2-38(b)是按 sex、age 两个变量值拆分的结果。读者可以用 data02-05a 做实验。

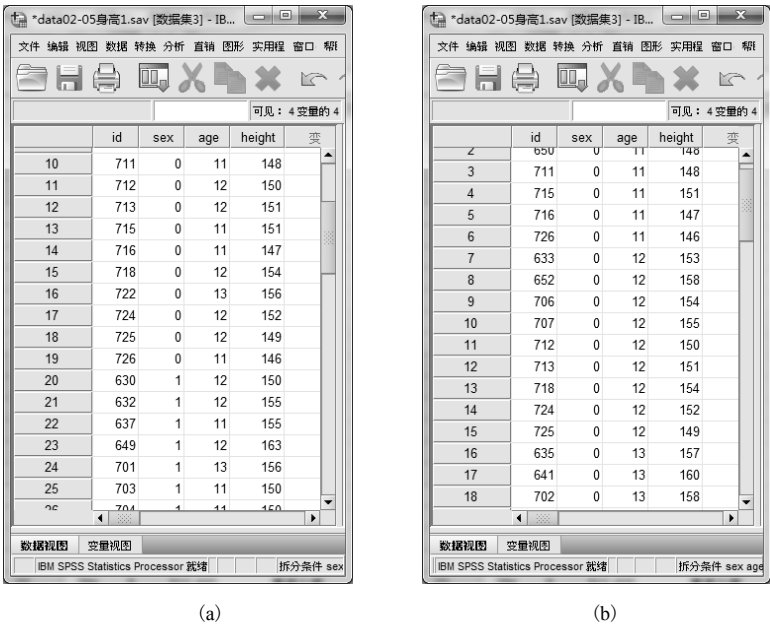


图 2-38 选取不同拆分变量的拆分结果

2. 合并数据文件

(1) 两种合并方式

合并数据文件是指将外部数据中的观测或者变量合并到当前数据文件中去，它包括两种合并方式。

① 从外部数据文件增加观测到当前数据文件中，这种方法称为纵向合并或追加观测。相互合并的数据文件中应该有相同的变量，不同的观测。

② 从外部数据文件增加变量到当前数据文件中，称为横向合并。相互合并的数据文件中包含不同的变量。

(2) 增加观测的纵向合并

① 首先在数据窗口中打开一个数据文件 data02-06，如图 2-39(a)所示。与一个未打开的数据文件 data02-07 合并。data02-07 的数据如图 2-39(b)所示。两个数据文件都有相同的变量 id、gender、age。

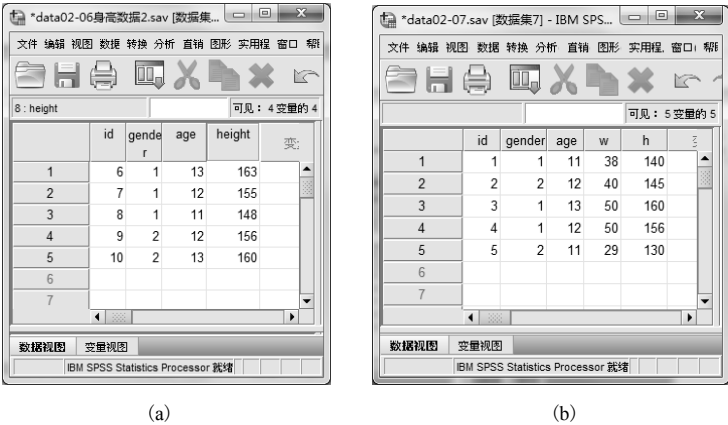


图 2-39 两个数据文件的原始状态

② 按【数据→合并文件→添加个案】顺序，打开【将个案添加到 data02-06.sav[数据集 6]】对话框，如图 2-40 所示。指定一个要与之合并的数据文件。有两种情况：

- 【打开的数据集】框中列出与 data02-06 同时打开的数据文件，可以从文件列表选择一个与之合并。
- 【外部 SPSS Statistics 数据文件】，指定一个未打开的 SPSS 数据文件与 data02-06 合并。单击【浏览】按钮，指定一个外部 SPSS 数据集。

指定与主文件合并的数据文件后，单击【继续】按钮，打开如图 2-41 所示的对话框。

③ 【新的活动数据集中的变量】框中列出的变量是在两个数据文件中变量名相同、类型相同的变量(id、sex、age)。这些变量直接包括在合并后的新文件中。

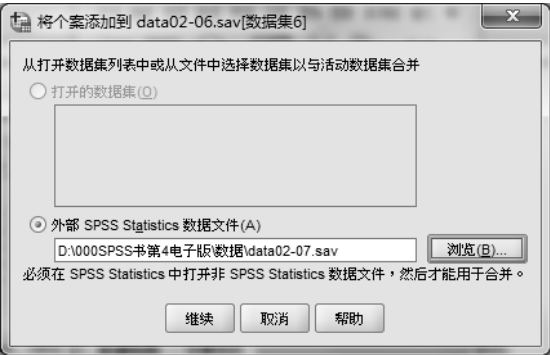


图 2-40 指定与主文件合并的数据文件



图 2-41 增加观测对话框

【非成对变量】框中列出的变量是未配对变量，有 height、h 和 w 这些在另一个数据文件中

找不到变量名和类型与之相同的变量，即它们不能配对。标有“\*”的是当前数据文件中的变量，标有“+”的是外部数据文件中的变量。

在对话框下方列出两个变量标识符号，各代表在哪个数据集中的变量。

- ④ 根据情况处理数据。
- 只合并两个数据文件中变量名和类型都相同的变量的观测时，单击【确定】按钮。
  - 追加外部数据文件中名称不同的变量(不匹配变量)的观测。此时需要首先在【非成对变量】框中设置配对变量，即先选取一个变量，再按住 Ctrl 键选取与之配对的变量，然后单击【对】按钮将它们送入新的数据文件变量表中，显示“height & h”，最后单击【确定】按钮。没有配对的变量，也可以出现在合并后的数据集中，只需选择并送入右面的框中。
  - 也可以改名后再送入右框。在非成对变量中选择一个变量，单击【重命名】按钮，在【重命名】对话框中给出新名，单击【继续】按钮即可。
  - 将个案源表示为变量。指定一个变量，值为 1 表明来自工作数据文件的观测，值为 0 表明是外部数据文件中的观测。默认的变量名为源 01，也可以自己命名。

【例 6】图 2-39(a)所示当前工作数据文件中的变量 height 与外部数据文件中的变量 h 如图 2-39(b)所示，均为身高数据，只是变量名称不同。在【非成对变量】框中选择这两个变量，单击【对】按钮。在新工作数据文件的变量表中显示“height & h”。选择【将个案源表示为变量】项，指定生成指针变量，使用默认名“源 01”。单击【确定】按钮，合并结果如图 2-42(a)所示。



	id	sex	age	height	源01
1	6	1	13	163	0
2	7	1	12	155	0
3	8	1	11	148	0
4	9	2	12	156	0
5	10	2	13	160	0
6	1	1	11	140	1
7	2	2	12	145	1
8	3	1	13	160	1
9	4	1	12	156	1
10	5	2	11	130	1
11					

(a)



	id	sex	age	height	w	源01
1	6	1	13	163	.	0
2	7	1	12	155	.	0
3	8	1	11	148	.	0
4	9	2	12	156	.	0
5	10	2	13	160	.	0
6	1	1	11	140	38	1
7	2	2	12	145	40	1
8	3	1	13	160	50	1
9	4	1	12	156	50	1
10	5	2	11	130	29	1
11						

(b)

图 2-42 不同变量情况的观测合并结果

未配对变量表中的变量在配对时要求必须具有相同的变量类型。宽度不相同，当前文件中的变量宽度应当大于等于外部文件变量的宽度(如 height 的宽度大于等于 h 的宽度)。如果当前文件中的变量宽度小于外部文件变量的数据的宽度(如果 height 的宽度小于 h 的宽度)，在合并后外部文件被合并的观测中的相应变量数据会丢失。若干个星号“\*”表示丢失的变量值。

对于只在一个数据文件中含有的变量(例如，变量 w 仅在外部文件中存在)，如果不进行配对，但要求包含在新的数据文件中，只要选择这个变量，并将其移入新数据文件变量表中即可。

图 2-40 (b) 是将  $w$  变量移入新的活动数据集中的变量, 但  $w$  变量并没有与之配对的变量, 由于当前工作数据文件不包括  $w$  变量, 因此相应的观测  $w$  值为缺失值。

3. 增加变量

增加变量有两种方式:

- 两个数据文件按观测顺序一对一地横向合并;
- 按关键变量合并, 即要求两个数据文件必须有一个共同的关键变量, 两个数据文件中关键变量值相同的观测合并为一个观测。

下面以 data02-08 为当前工作数据文件, 包括变量 id、sex、age、h、w, data02-09 为外部数据文件, 包括变量 id、w。以这两个数据文件横向合并为例, 说明操作步骤。

(1) 打开数据文件 data02-09, 显示在另一个数据编辑窗口中。

(2) 在数据文件 data02-08 编辑窗口中, 按【数据→合并文件→添加变量】顺序, 打开【将变量添加到 data02-08.sav[数据集 3]】对话框, 见图 2-43。在【打开的数据集】栏内显示已经打开的数据文件 data02-09。单击选择这个文件, 然后单击【继续】按钮。

(3) 在打开的(如图 2-44 所示)【添加变量从数据集 2】对话框中, 左栏已排除的变量列出的是两个文件中的同名变量, 只有这样的变量可以作为关键变量。对话框右侧【新的活动数据集】矩形框中, 列出了可以在新工作数据文件中存在的变量。

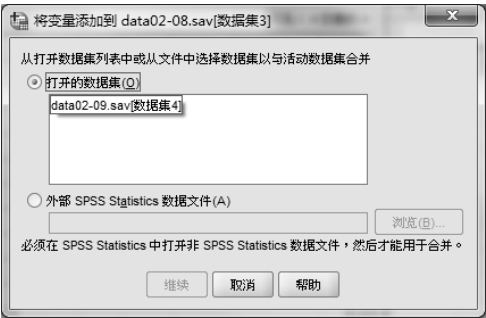


图 2-43 【将变量添加到 data02-08.sav[数据集 3]】对话框

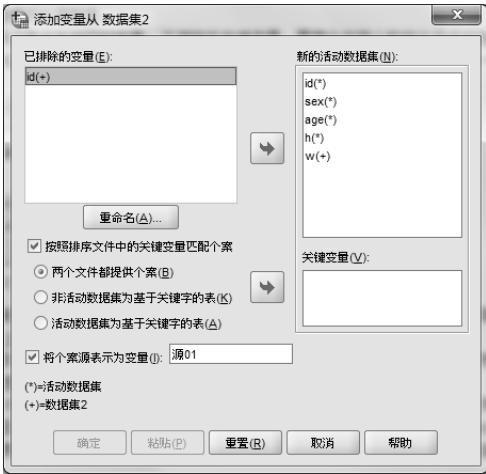


图 2-44 【添加变量从数据集 2】对话框

在两个矩形框中标有“\*”的是当前工作数据文件中的变量, 本例中是打开的工作数据文件 data02-08.sav; 标有“+”的是指定的外部数据文件或已经打开的另一个数据文件中的变量, 本例中是已经打开的数据文件 data02-09.sav 中的变量。

(4) 根据情况处理数据。

- ① 如果没有名字相同的变量, 则不用指定关键变量。要想合并两个数据文件中的变量, 单击【确定】按钮即可开始横向合并两个数据文件了。结果是按观测出现的顺序一对一地合并。
- ② 如果两个数据文件中有同名的变量, 那么合并的结果保留当前数据文件中同名的变量加上外部数据文件中不同名的变量。
- ③ 选择在当前数据文件与外部数据文件中包含的同名变量作为关键变量, 需要先对数据文件按关键变量值的升序排序。

将排序后关键变量值相同的合并为一个观测。图 2-45 (a) 为当前数据文件 data02-08，图 2-45 (b) 为第 2 个数据文件 data02-09，均已经按关键变量 id 排序。图 2-45 (c) 为合并后的数据文件。观测 id=60, 65, 68 的观测，在两个数据文件中都存在，横向合并。

对于两个文件中关键变量值不同的观测，处理方法是，选择按照排序文件中的关键变量匹配个案。激活下面 3 个选项，在其中选择一种处理方式。

- **【两个文件都提供个案】**。即观测由两个数据文件提供。合并的结果是将第 2 个数据文件的观测追加到当前工作数据文件中，如图 2-45 (c) 中 id=60, 64, 65, 67、68 的观测。与 data02-09 的 id 值相同的 id=60, 65, 68 合并, id 值不相同的 id=64, 67 也追加到 data02-08 文件中，结果保存在 data02-08a.sav 中。
- **【非活动数据集为基于关键字的表】**。即保持当前数据文件中的观测数目不变。在第 2 个数据文件中，只有那些与当前数据文件中关键变量等值的观测才能合并到工作数据文件中，例如，数据文件 data02-08 与 data02-09 以这种方式合并，其结果如图 2-46 (a) 所示。数据保存在 data02-08b.sav 中。
- **【活动数据集为基于关键字的表】**。当前数据文件中的观测与第 2 个文件中的关键变量值相等时并入第 2 个文件，例如，数据文件 data02-08 与 data02-09 以这种方式合并，其结果如图 2-46 (b) 所示。

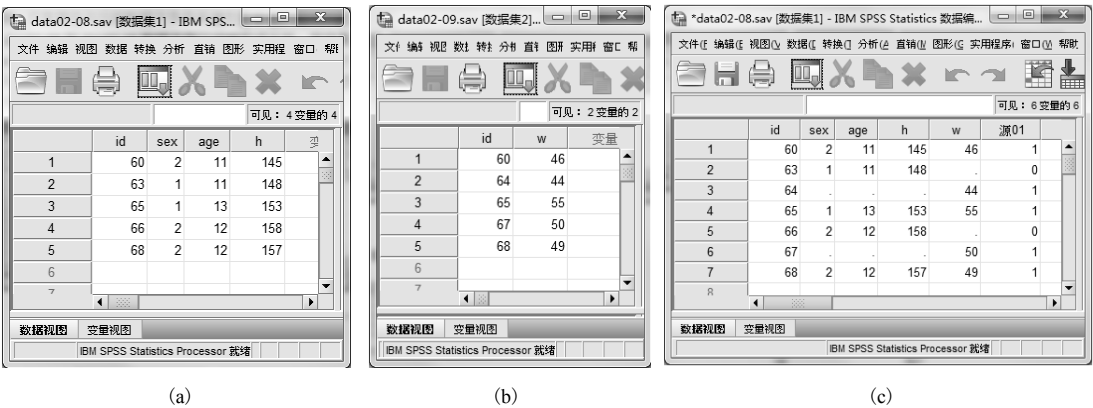


图 2-45 由两个排序数据文件提供合并数据

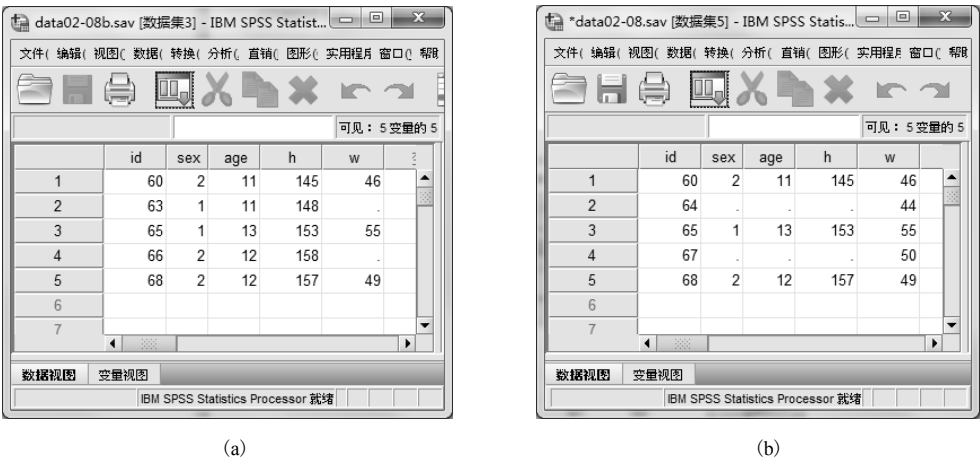




图 2-46 以关键变量值相等的原则合并

最后将在【已排除的变量】框中选择的关键变量(id)，通过单击下面一个按钮移到【关键变量】框中。单击【确定】按钮将指定条件和方式的合并提交系统执行。系统将提示警告：如果两个文件没有按关键变量排序，合并可能失败。因此，在执行合并功能之前，必须将内、外两个文件均按关键变量排序。

- (5) 几点说明。
- ① 如果在当前数据文件中与外部数据文件中有同名的变量，则外部数据文件中的变量列于【已排除的变量】框中，当前数据文件中的变量列于右面【新的活动数据集】框中。
  - ② 【已排除变量】框中的变量若选为关键变量，则可以将其移到【关键变量】框中。与其同名的【新的活动数据集】框中的同名变量消失。
  - ③ 如果一定要将【已排除变量】框中，外部数据的同名变量合并到新的数据文件中去，那么应先为该变量更名；即单击【重命名】按钮，在被打开的相应对话框中赋予该变量一个新名。然后选择该变量，并单击上面一个按钮将其移到【新的活动数据集】框中。
  - ④ 【新的活动数据集】框中的变量均为新数据文件中的变量。如果不想使某变量出现在框中，则选择这个变量，将其移到左面【已排除的变量】框中。
  - ⑤ 为变量更名。如果两个数据文件中有同名变量，但内容不同，需要对其中一个变量更名；如果两个文件中作为关键变量的两个变量不同名，则应该改成相同的变量名。
  - ⑥ 生成新变量。如果选中【将个案源表示为变量】。即显示数据来源变量，一个新的变量(读者输入的变量名称)将会加入到当前数据文件中。其变量值 0 表示观测来自当前数据文件，1 表示观测来自非工作数据文件。

2.3.2 观测的排序与排秩

1. 观测排序

在进行数据处理过程中，有时需要按照某个或某些变量(排序变量)值的大小重新排列观测在数据文件中出现的先后顺序，可按下述步骤实现。



图 2-47 【排序个案】对话框

中，按第二观测的值排序，依此类推。

如果排序变量是字符型的，则英文排序按拼写的字母 ASCII 码顺序排列，中文排序按拼音字母的 ASCII 码顺序排列。

- (1) 按【数据→排序个案】顺序打开【排序个案】对话框，如图 2-47 所示。
  - (2) 在左侧的源变量框中选择排序变量，移到右面的【排序依据】框中。
- 如果选择了两个以上的排序变量，则列于首位的称为第一排序变量，其后的顺序分别称为第二排序变量、第三排序变量……排序的结果与排序变量在【排序依据】框中的顺序有关。
- 排序的结果是观测先按第一排序变量的值排列观测，在第一排序变量值相等的观测组



(3) 确定排序的方式，即根据变量顺序进行排列。

① 在【排序依据】框内选择一个排序变量。

② 在【排列顺序】栏内选择以下一种排序方式：

- 【升序】。按所选择的排序变量值的升序排列；
- 【降序】。按所选择的排序变量值的降序排列。

(4) 重复第3步的操作，可以指定下一个排序变量的排序方式。

(5) 单击【确定】按钮，即可完成排序工作。单击【粘贴】按钮可以在语句窗口中生成程序语句。

## 2. 根据变量的值对观测排秩

在当前数据文件中产生秩变量的操作步骤如下：

(1) 按【转换→个案排秩】顺序单击菜单项，打开【个案排秩】对话框，如图 2-48 所示。

(2) 在左侧的源变量框中选择至少一个变量进入右侧的【变量】框中。对每个变量产生一个秩变量。

(3) 在【将秩 1 指定给】栏中选择秩的排列方式：

- 【最小值】。定义 1 为最小数值的秩；
- 【最大值】。定义 1 为最大数值的秩。

(4) 读者可以选择一个或多个分组变量进入【排序标准】框中，系统将按排序标准变量的值分组排秩。

图 2-46 中选择了身高变量为排秩的对象，即对身高排秩；选择了性别变量作为排秩的分组变量。

(5) 单击【秩的类型】按钮，打开如图 2-49 所示的对话框，指定产生的秩的算法。

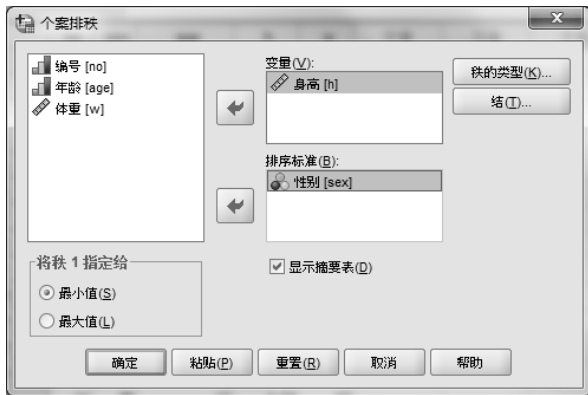


图 2-48 【个案排秩】对话框



图 2-49 【个案排秩: 类型】对话框

① 【秩】。简单秩。数据文件中的新变量值就是对应的观测的秩，这是默认方式。新变量名为原变量名前冠以“r”。

② 【Savage 得分】。秩变量的值是依据指数分布所得的 Savage 分数。新变量名为原变量名前冠以“s”。

③ 【分数秩】。新变量的值等于简单秩除以非缺失观测的加权和。

④ 【%分数秩】。秩值为其简单秩除以所有具有合法值的观测数目乘以 100。

⑤ 【个案权重总和】。新变量的值是观测权重之和。在同组中新变量值是个常数。

⑥【Ntiles】，分段排序。在参数框中输入分段数，分段数必须是大于 1 的整数。某一观测的秩值是按该观测占的百分位数的位置来决定的。例如，如果输入的数值为 4，那么变量值的百分位数低于 25%的观测的秩将被赋值 1，位于 25%~50%的观测的秩将被赋值 2，位于 50%~75%的观测的秩将被赋值 3，高于 75%的观测的秩被赋值 4。

⑦【比例估计】。是与一个特定秩的分布的累计比估计。

⑧【正态得分】。即与估计累计比相应的 Z 分数。

选择了⑦、⑧秩类型后，激活下面的选项，可以进一步指定计算新的秩变量值的公式，即在【比例估计公式】栏中进行选择。

- 【Blom】，由公式  $(r-3/8)/(w+1/4)$  决定，其中  $r$  为秩， $w$  为观测的权重之和。此项为默认设置。
- 【Tukey】，由公式  $(r-1/3)/(w+1/3)$  决定，其中  $w$  为观测权重之和， $r$  为秩。
- 【Rankit】，由公式  $(r-1/2)/w$  决定， $w$  为观测权重之和， $r$  为秩，范围为  $1\sim w$ 。
- 【Van der Waerden】选项，由公式  $r/(w+1)$  决定，此处  $w$  为观测权重之和， $r$  为秩，范围为  $1\sim w$ 。

若要求这是默认设置，可单击此项，输出窗口不显示这些信息。

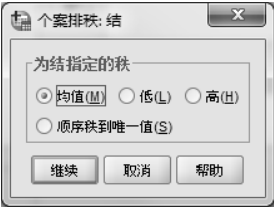


图 2-50 【个案排序：结】对话框

(6) 确定结的秩。

变量值相同的称为结。结的秩次的决定原则可以在【个案排序：结】对话框中指定。在如图 2-48 所示的主对话框中单击【结】按钮，打开【个案排序：结】对话框，如图 2-50 所示。

- ①【均值】。相同值的秩取平均值。
- ②【低】。相同值的秩取最小值。
- ③【高】。相同值的秩取最大值。

④【顺序秩到唯一值】。相同值的秩取第一个出现的秩次值，其他观测秩次顺序排列。体育比赛常用这种方法排列名次，即并列第\*名。

上述 4 种方法排序的比较如表 2-4 所示。

表 2-4 4 种方法排序的比较

观 测 值	均 值	低	高	顺序秩到唯一值
1.00	1	1	1	1
2.50	3	2	4	2
2.50	3	2	4	2
2.50	3	2	4	2
3.50	5	5	5	3
3.75	6	6	6	4

(7) 以上选项确定后，单击【确定】按钮，根据指定的变量、分组变量及其他选项计算秩，并生成新变量。在输出窗口中显示新变量的名称、标签、秩类型等总结性的信息。

2.3.3 对变量值重新编码

把连续变量变成分类变量或重新分类时需要重新编码。【转换】菜单中的【重新编码】命令和【自动重新编码】命令可以对多个类型相同的变量重新编码，生成新变量。新变量的值是重新编码的结果，也可以用新代码代替原始变量。重新编码命令允许在编码过程中进行人为干预。

1. 使用重新编码命令重新编码

【转换】菜单中有两项与重新编码有关：

- 【重新编码为相同变量】。对一个变量重新编码，结果代替该变量，主对话框如图 2-51 所示。
- 【重新编码为不同变量】。生成新变量，变量的值是编码的结果，主对话框如图 2-52 所示。

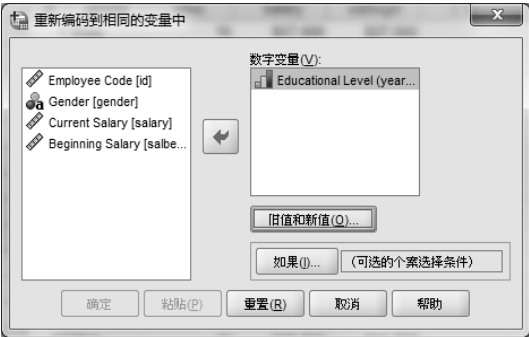


图 2-51 【重新编码到相同的变量中】主对话框



图 2-52 【重新编码为其他变量】主对话框

两个选项的主对话框区别仅在于【重新编码到相同的变量中】对话框中没有定义输出变量的部分。因此，在此只叙述【重新编码为其他变量】生成新变量的操作。以数据文件 data02-10 为例说明对年龄 age 变量重新编码方法。

- (1) 从变量列表中选择要重新编码的变量，送入【数字变量→输出变量】框中。
- (2) 每选择一个变量，就在【输出变量】的【名称】栏内输入新变量名，在【标签】栏内输入新变量标签。单击【更改】按钮。
- (3) 可以单击【如果】按钮打开相应对话框，根据条件选择要编码的观测。

单击【旧值和新值】按钮，打开【重新编码到其他变量：旧值和新值】对话框，如图 2-53 所示。左面【旧值】给出原变量值或值范围的区域，每选择一项就在右边的【新值】栏中选择一项，或选择一项同时给出新变量的值。单击【添加】按钮，将新、旧变量之间关系，即变量值与编码的对应关系送入【旧→新】栏中。

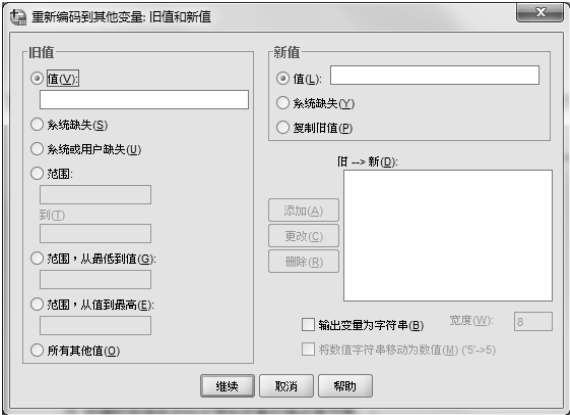


图 2-53 【重新编码到其他变量：旧值和新值】对话框

单击【添加】按钮，将新、旧变量之间关系，即变量值与编码的对应关系送入【旧→新】栏中。

- ① 【旧值】栏中选择并给出原始变量的值或值范围。
  - 【值】栏中输入单个值。
  - 【系统缺失】。为系统缺失值。
  - 【系统或用户缺失】。为系统缺失值或用户缺失值。
  - 【范围】：【\_\_到\_\_】框，在两个输入区中给出最低和最高两个值，定义这个区间内的所有值。
  - 【范围，从最低到值】框，给出一个值，定义小于等于这个值范围内的值。
  - 【范围，从值到最高】框，给出一个值，定义大于等于这个值范围内的值。

- 【所有其他值】。定义前面所有定义没有包括的值。
- ② 【新值】栏中，针对【旧值】栏中给出的值，给出新代码。
- 在【值】框针对旧值给出的值，输入对应的新代码的值。
- 【系统缺失】。将旧值给出的值定义为缺失值。
- 【复制旧值】。新代码与旧值给出的值相同。

注意：各段值的衔接，表达式一定不要漏掉某些介于两组值之间的值。  
如果要按不同情况分组定义，还应该单击【如果】按钮进入对话框，阐明条件。有关操作见本章 2.1.5 节“根据已有的变量建立新变量”。

(4) 定义结束，单击【继续】按钮，返回主对话框，单击【确定】按钮，开始转换。

【例 7】 重新编码实例。

打开数据文件 data02-11。

(1) 编码要求：对职工的起始工资 salbegin 和当前工资 salary 重新编码，对应的新变量名分别为 salbegin1 和 salary1，将连续变量编码为分类变量。要求的代码如表 2-5 所示。

(2) 操作要点

① 按【转换→重新编码到不同变量】顺序单击菜单项，打开对话框，将 salary 和 salbegin 送入【数字变量→输出变量】框中。

单击 salary->?, 在【输出变量】栏的【名称】后输入新变量名 salary1，在【标签】后输入标签“当前工资等级”，单击【更改】按钮，则在【数字变量→输出变量】框中显示 salary→salary1。

表 2-5 编码表

salary 和 salbegin	<=16000	16001~20000	20001~25000	25001~30000	30001~35000	35001~40000	40001~45000	45001~50000	50001~55000	55001~60000	60001~65000	>=65001	System-missing
salary1 和 salbegin1	1	2	3	4	5	6	7	8	9	10	11	12	System-missing

再单击 salbegin，在【输出变量】栏的【名称】后输入新变量名 salbegin1，在【标签】后输入标签“起始工资等级”，单击【更改】按钮。【数字变量→输出变量】框中显示 salbegin→salbegin1。

单击【旧值和新值】按钮，打开相应对话框，定义新旧变量对应关系。

② 【旧值】栏中选择【范围，从最低到值】，输入 16000，在【新值】栏的【值】中输入 1。单击【添加】按钮，【旧→新】框中显示 lowest thru 16000→1。

③ 【旧值】栏中选择【范围，\_\_到\_\_】，输入值 16001 和 20000，在【新值】栏的【值】中输入 2。单击【添加】按钮。【旧→新】栏内显示 16001 thru 20000→2；新变量代码为 3~11 的都与此操作相同。

④ 【旧值】栏中选择【范围，从值到最高】，输入 65001，在【新值】栏的【值】中输入 12。

⑤ 【旧值】栏中选择【系统缺失】，在【新值】栏也选择【系统缺失】。单击【添加】按钮，【旧→新】框中显示 SYSMIS→SYSMIS。

定义完成，单击【继续】按钮，返回主对话框，单击【确定】按钮，提交运行，结果见数据文件 data02-11a.sav。数据编辑器的数据视图如图 2-54 所示。

(3) 对于等级较多的重新编码，写个小程序会更简单。这个程序在编码完成后的输出窗中也可以看到。本例运行程序如下：

```
RECODE salary salbegin
  (SYSMIS=SYSMIS) (Lowest thru 16000=1) (16101 thru 20000=2)
  (20001 thru 25000=3) (25001 thru 30000=4) (30001 thru 35000=5)
  (35001 thru 40000=6) (40001 thru 45000=7) (45001 thru 50000=8)
  (50001 thru 55000=9) (55001 thru 60000=10) (60001 thru 65000=11)
  (65001 thru Highest=12) INTO salary1 salbegin1 .
VARIABLE LABELS salary1 '工资等级' /salbegin1 '初始工资等级'.
EXECUTE .
```

程序由两部分组成。

第一段程序是 RECODE 过程语句。RECODE 是命令关键字，后面是要进行重新编码的原始变量。中间部分是编码规则的表达式。最后 INTO 跟着两个新变量名。

每个编码表达式由等号连接两部分，等号前是原始变量值或值范围表达式，等号后面是新代码值。表达式有几种形式：(SYSMIS=SYSMIS)表示新变量的系统缺失值与原始变量定义相同；(Lowest thru C1=C3)定义原始变量值小于等于 C1 的，新变量的值为 C3；(C1 thru C2=C3)定义原始变量值在 C1 与 C2 之间的，包括 C1、C2，新变量的值为 C3；(C1 thru Highest=C3)定义原始变量的值大于等于 C1 的，新变量的值为 C3。

第二段程序是 VARIABLE LABELS 过程语句，为新变量加变量标签。VARIABLE LABELS 是过程语句关键字，后边是变量名与变量标签，中间用空格分隔。

2. 自动重新编码

(1) 使用自动编码功能对数据进行重新编码是对数据进行预处理的需要。

① 原始分类变量的分类值不是等间隔的，会在进行频数分布分析时形成空单元，不但浪费计算机资源，也使输出表格臃肿，不利于得出结论。

② 有些分析过程要求参与分析的分类变量必须是数值型的，不能是字符型的，需要转换。某些分析过程要求分类变量值是整数。

(2) 以数据文件 data02-11 中的受教育程度变量 educ 为例，说明自动重新编码的操作。

① 按【转换→自动重新编码】顺序单击菜单项，打开相应的对话框，如图 2-55 所示。

将要自动编码的变量 educ 送入右面的【变量→新名称】栏，显示 educ→educ 1，在下面的【新名称】栏输入新变量名 educ1，单击【添加新名称】按钮，【变量→新名称】栏中显示新旧变量名对应关系：educ→educ1。



图 2-54 重新编码后的数据（两个新变量）



图 2-55 【自动重新编码】对话框

② 在【重新编码的起点】栏中选择【最低值】，表示从最小值开始编码。也可以选择从最大值开始编码，但是对受教育程度这个有序分类变量来说，最好新编码顺序与原来的受教育年限的原始值一致。单击【确定】按钮。在输出窗口显示编码结果，如图 2-56(a)所示。

③ 输出结果表明，原始值从 8~21，缺少 9、10、11、13，新变量值从 1~10，使用原始值作为值标签。如果在分析输出表中使用值标签，就可以得到比较满意的、易于解释的结果。数据编辑器的数据视图如图 2-56(b)所示。

educ into educ1 (Educational Level (years))

Old Value	New Value	Value Label
8	1	8
12	2	12
14	3	14
15	4	15
16	5	16
17	6	17
18	7	18
19	8	19
20	9	20
21	10	21

(a)

	id	gender	educ	salary	salbegin	educ1	变量
1	1	Male	15	\$57,000	\$27,000	15	
2	2	Male	16	\$40,200	\$18,750	16	
3	3	Female	12	\$21,450	\$12,000	12	
4	4	Female	8	\$21,900	\$13,200	8	
5	5	Male	15	\$45,000	\$21,000	15	
6	6	Male	15	\$32,100	\$13,500	15	
7	7	Male	15	\$36,000	\$18,750	15	
8	8	Female	12	\$21,900	\$9,750	12	
9	9	Female	15	\$27,900	\$12,750	15	
10	10	Female	12	\$24,000	\$13,500	12	
11	11	Female	16	\$30,300	\$16,500	16	
12	12	Male	8	\$28,350	\$12,000	8	
13	13	Male	15	\$27,750	\$14,250	15	
14	14	Female	15	\$35,100	\$16,800	15	
15	15	Male	12	\$27,300	\$13,500	12	

(b)

图 2-56 自动重新编码的结果

(3) 对话框中还有几个选项：

- ① 【对所有变量使用相同的重新编码设计】。
- ② 【把空字符串值视为用户缺失值】。
- ③ 【模板】选项。
  - 【从文件应用模板】：选择此项，指定一个模板文件作为设置本数据文件指定变量重新编码的模板。单击该选项后面的【文件】按钮，指定模板文件。
  - 【将模板另存为】：把当前的重新编码方案作为模板保存起来，以便以后用在其他变量的重新自动编码上。选择此项，单击【文件】按钮，指定保存文件的位置和文件名。


2.3.4 数据文件的转置与重新构建

分析工作中要求的数据排列方式往往与当前数据文件中的数据排列方式不同。为了满足分析过程对数据文件结构的要求，就需要进行变换。使用移动、复制固然可以达到目的，但是往往容易出错。本节介绍由变换工具自动变换的方法。

1. 数据文件的转置

利用【数据】的【转置】功能，可以将数据文件中原来的行变成列，原来的列变成行；将观测转变为变量，将变量转变为观测；在新文件中建立一个其值为原来变量名的变量。转置后的数据文件与原来的数据文件完全不同，应该保存到另一个文件名下。

操作步骤如下：

- ① 按【数据→转置】顺序打开【转置】对话框，如图 2-57 所示。
- ② 在左侧的源变量框中选择要进行转置的变量，单击  按钮送到【变量】框中。

这些变量在新数据文件中变成观测(从列变成行)。新数据文件不会出现未选择的变量。

③ 从源变量框中选择一个变量送入【名称变量】框中。该变量的值在新数据文件中作为变量名出现。一般选择标识观测的变量，如观测的编号、姓名等。如果它是一个数值变量，则新变量名为该变量各值冠以字母“K\_”。如果不选择【名称变量】，系统会自动给转置后的变量赋予名称 var001、var002、…、var00n。

- ④ 单击【确定】按钮，进行转置。

【例 8】 数据文件转置操作实例。

数据文件 data02-12.sav 是对汽车市场调查的数据——25 名被访者对凯迪拉克、雪佛龙等 17 个品牌的汽车打分的结果数据，18 个变量为 17 个汽车名称和一个被访者编号。每个观测就是一个被访者给 17 种车的打分。为了进行顾客偏好分析，需要对数据文件进行转置。转置结果经整理保存到 data02-12a 中。以此数据为例，说明转置操作。

本例选择 17 种品牌变量作为要转置的变量送入【变量】框中。本例选择被访者编号变量 number 送入【名称变量】框中。

图 2-58 (a) 为转置前的变量视图窗口，变量是编号和 17 个品牌的汽车；图 2-58 (b) 为转置后的变量视图窗口，变量为 K\_1~K\_25 和系统自动生成的变量 CASE\_LBL。

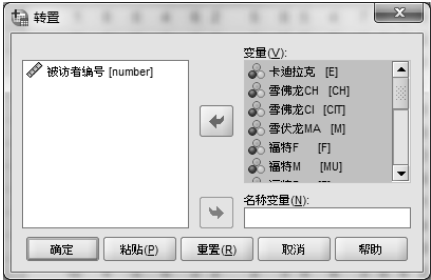


图 2-57 【转置】对话框

	名称	类型	宽度	小数	标签
1	number	数值(N)	3	0	被访者编号
2	E	数值(N)	8	0	凯迪拉克
3	CH	数值(N)	8	0	雪佛龙CH
4	CI	数值(N)	8	0	雪佛龙CI
5	M	数值(N)	8	0	雪佛龙MA
6	F	数值(N)	8	0	福特F
7	MU	数值(N)	8	0	福特M
8	P	数值(N)	8	0	福特P
9	A	数值(N)	8	0	本田A
10	CI	数值(N)	8	0	本田C
11	CO	数值(N)	8	0	林肯C
12	G	数值(N)	8	0	普利茅斯G
13	H	数值(N)	8	0	普利茅斯H
14	V	数值(N)	8	0	普利茅斯V
15	FI	数值(N)	8	0	庞蒂阿克
16	D	数值(N)	8	0	大众D
17	R	数值(N)	8	0	大众R
18	DL	数值(N)	8	0	沃尔沃D

(a)

	名称	类型	宽度	小数
1	CASE_LBL	字符串	3	0
2	K_1	数值(N)	8	2
3	K_2	数值(N)	8	2
4	K_3	数值(N)	8	2
5	K_4	数值(N)	8	2
6	K_5	数值(N)	8	2
7	K_6	数值(N)	8	2
8	K_7	数值(N)	8	2
9	K_8	数值(N)	8	2
10	K_9	数值(N)	8	2
11	K_10	数值(N)	8	2
12	K_11	数值(N)	8	2
13	K_12	数值(N)	8	2
14	K_13	数值(N)	8	2
15	K_14	数值(N)	8	2
16	K_15	数值(N)	8	2
17	K_16	数值(N)	8	2
18	K_17	数值(N)	8	2
19	K_18	数值(N)	8	2
20	K_19	数值(N)	8	2
21	K_20	数值(N)	8	2
22	K_21	数值(N)	8	2
23	K_22	数值(N)	8	2
24	K_23	数值(N)	8	2
25	K_24	数值(N)	8	2
26	K_25	数值(N)	8	2

(b)

图 2-58 转置前后的变量视图

在输出窗口中列出了所有新变量名。另外，如果数据包含缺失值，那么 SPSS 将其设置为系统的缺失值。为了保留缺失值，可以重新改变对变量中缺失值的设置。图 2-59 为转置前的数据视图窗口，图 2-60 为转置后的数据视图窗口。原来的 17 个变量转换成 17 个观测。

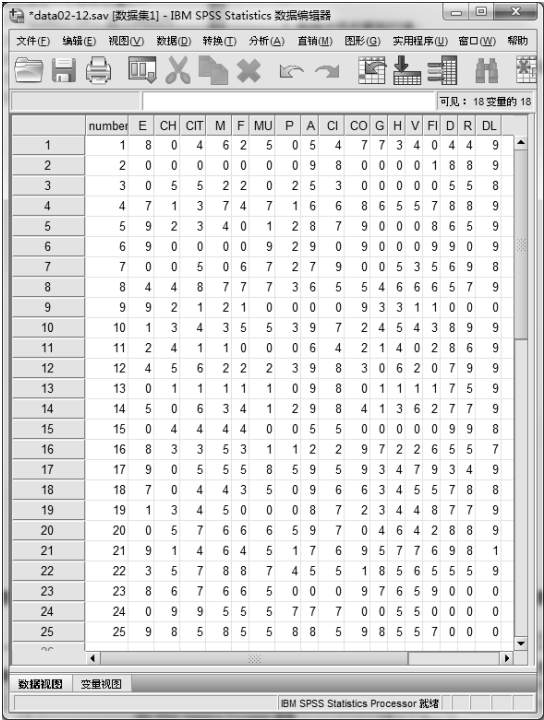


图 2-59 转置前的数据视图窗口

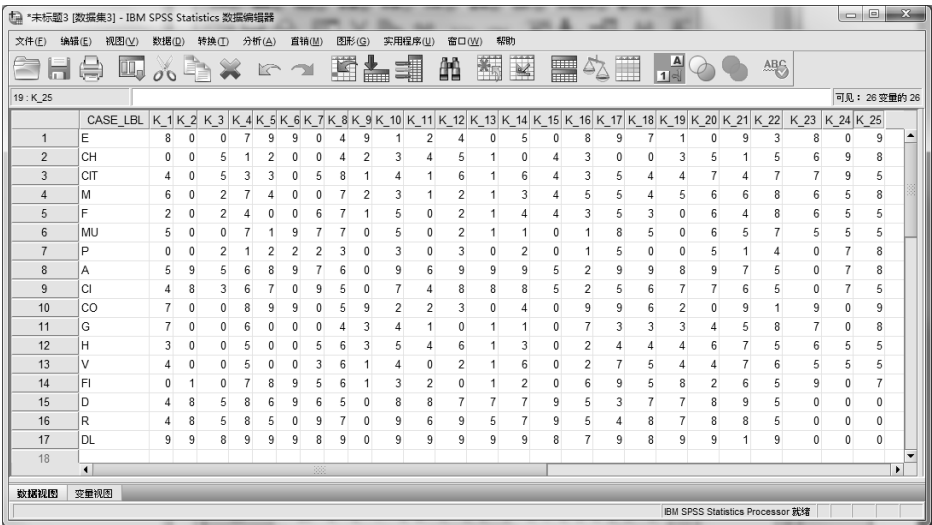


图 2-60 转置后的数据视图窗口

2. 数据文件的重新构建

(1) SPSS 数据文件结构

SPSS 数据分析所需要的数据文件在【数据观察窗口】中的结构分为 3 种。



① 简单数据文件。一个变量占一列，一个观测占一行。例如，对一个班的所有学生进行一项测试，所有分数仅出现在一列中，每个学生占一行。

② 有关一个观测的信息占不止一行。例如，一个因素的每个水平占一行或不止一行，如表 2-6 所示。一个因素的若干水平称作一个观测组。在 SPSS 数据分析中，当数据用这种方式构造时，因素经常作为分组变量。

③ 有关一个变量的信息占不止一列。例如，一个因素的每个水平占一列，如表 2-7 所示。一个因素的若干列称作一个变量组。在 SPSS 数据分析中，当数据按这种方式构造时，因素常常涉及重复测量。

表 2-6 观测组结构

factor	var
1	3
1	8
1	6
2	5
2	9
2	4

表 2-7 变量组结构

var1	var2
4	6
8	5
7	9

(2) 各种分析方法所需要的数据文件结构

① 要求观测组数据结构的分析过程。数据必须按观测组构建，以便做分组变量的分析，例如，一般线性模型中的单因变量方差分析、多因变量方差分析、方差成分分析；混合模型；OLAP Cubes 和独立样本 T 检验或非参检验。

② 要求变量组数据结构的分析过程。数据必须按变量组构建的分析有：一般线性模型的重复测量，Cox 回归分析中时间为因变量的协方差分析、配对样本 T 检验或相关样本的非参检验。

如果选择的分析过程所需要分析的数据结构与当前数据文件中的结构不符，需要进行变换，这项工作可以由【数据】菜单中的【重组】功能来完成。

【例 9】 变量组结构到观测的转换步骤。

要重组的数据文件为 data02-13-1，如图 2-61 所示。

① 单击【数据→重组...】打开【重组数据向导】对话框，如图 2-62 所示。这是一个向导式的操作，主对话框中有 3 个选项，对应 3 种重新构建的类型。

- 【将选定变量重组为个案】。如果当前数据文件的结构是变量组的，要转换成观测组结构，则选择此项。
- 【将选定个案重组为变量】。如果当前数据文件的结构是观测组的，要转换成变量组结构，则选择此项。
- 【转置所有数据】。如果选择此项，则单击【完成】按钮，自动关闭如图 2-62 所示的对话框，打开如图 2-57 所示的【转置】对话框，进行转置操作。

② 一个变量组的结构转换成观测组结构。以数据文件 data02-13-1 为例，5 个学生，学号分别为 1~5；A、B、C 三门课程的考试分数变量：scoreA、scoreB、scoreC，以及身高 h、体重 w 和学号 no 的记录，如图 2-62 所示。在【重组数据向导】主对话框中，【您希望做什么?】下面选择第一项，【将选定变量重组为个案】。单击【下一步】按钮，打开如图 2-63 所示的【重组

数据向导-第 2 步(共 7 步)】对话框。在对话框中选择【一个】，即要把一个变量组 scoreA、scoreB、scoreC 转换成观测组。如果有两组变量要同时进行转换，则应该选择第二项【多个】。

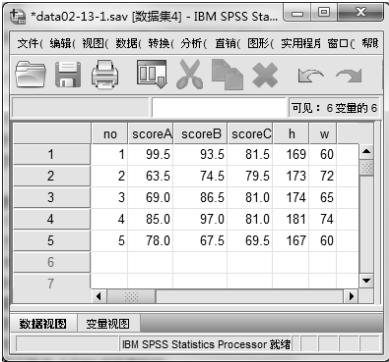


图 2-61 原始数据集

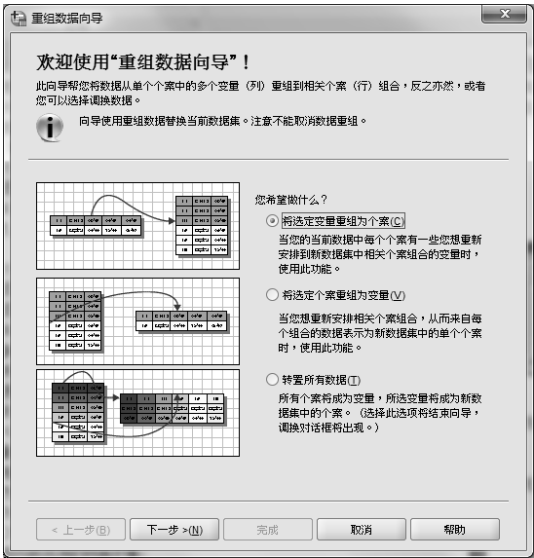


图 2-62 【重组数据向导】主对话框

③ 单击【下一步】按钮，打开如图 2-64 所示的【重组数据向导-第 3 步(共 7 步)】对话框。【个案组标识】栏，要求确定在新数据文件中的标识变量，其下拉列表中有 3 个选项：

- 【使用个案号】。
- 【使用选定变量】。
- 【无】。不用标识变量。

本例将【学号】作为观测标识，所以选择【使用选定变量】，在【当前文件中的变量】栏中选择变量【学号 no】送入右面的【变量】栏中。

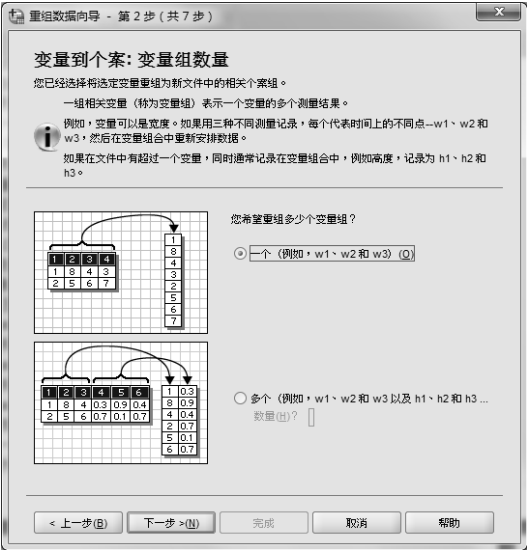


图 2-63 【重组数据向导-第 2 步(共 7 步)】对话框



图 2-64 【重组数据向导-第 3 步(共 7 步)】对话框

【要转置的变量】栏。在【当前文件中的变量】栏内选择要转换的变量并送入【目标变量】下面的栏中，本例选择 scoreA、scoreB、scoreC。在【目标变量】右面输入新文件中的变量名 score。

在【当前文件中的变量】栏中选择不进行转换，但还要出现在新文件中的变量，送入【固定变量】栏中。

本例还有身高 h、体重 w 两个变量在分析中不会用到，不希望出现在转换后的文件中，所以这项没有操作。

④ 单击【下一步】按钮，打开如图 2-65 所示的对话框。这一步，决定是否在新文件中生成索引变量。索引变量根据原始变量组，在新文件中按顺序编码。

【您希望创建多少索引变量？】。在 SPSS 过程中，索引变量可以作为分组变量，其下有 3 个选项：

- 【一个】。在大多数情况下，一个索引变量就足够了。本例选择此项。
- 【多个】。如果在当前文件中的变量组表现了多个因素的水平，可能要多个索引变量，输入想要产生索引变量的数目。
- 【无】。如果不需要生成索引变量则选择这一项。

所指定的数目会对下一步有影响，在下一步向导自动生成指定数目的索引变量。

⑤ 单击【下一步】按钮，打开如图 2-66 所示的【重组数据向导-第 5 步(共 7 步)】对话框。在这一步，确定索引变量的值。在【索引值是什么类型?】栏中有两个选项：

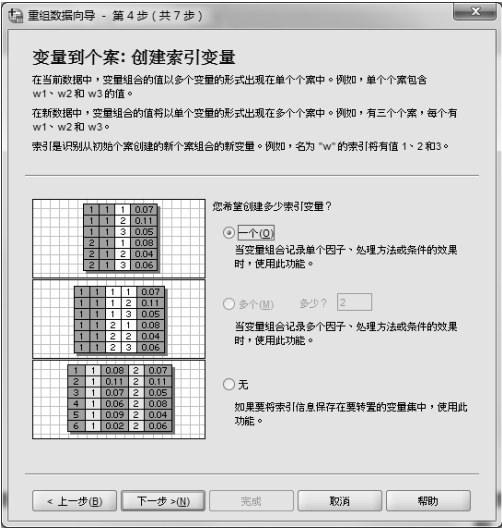


图 2-65 【重组数据向导-第 4 步(共 7 步)】对话框



图 2-66 【重组数据向导-第 5 步(共 7 步)】对话框

- 【有序数】。自动赋予顺序数作为索引值。若本例选择此项，则索引变量值为 1~3。
- 【变量名】。使用所选择的变量组各变量的变量名作为索引值。从列表选择一个变量组。本例若选择此项，则索引变量值为 scoreA、scoreB、scoreC。

【编辑索引变量名称和标签】。表中编辑索引变量属性，即对索引变量，可以改变其默认的变量名及输入描述变量的标签。

⑥ 单击【下一步】按钮，打开【重组数据向导-第 6 步(共 7 步)】对话框，如图 2-67 所示。

- 【处理未选定的变量】栏。处理原始数据文件中，未被选择的变量。在选择变量的第 3 步，选择了要重新构建的变量组和一个当前数据中的标识变量。所选择变量的数据将

出现在新文件中。如果在当前文件中还有其他变量，可以选择丢弃或保留它们。有两个单选项。

【从新数据文件中去掉变量】。丢掉未被选择的变量，系统默认。

【作为固定变量保持和处理】。

- 【所有已转置变量中的缺失值或空白值】栏，此栏中确定如何处理无效值，即要进行转换的变量中的缺失值和空值，有两个单选项：

【在新文件中创建个案】。

【废弃数据】。即从文件中剔除这个数据。

- 【个案计数变量】栏。确定是否在转换后的新文件中生成计数变量。计数变量包含当前数据中产生的新行数。如果选择丢弃无效值，计数变量可能是很有用的，因为有可能对给定的当前数据产生不同的行数，只有一个选项。

【计算由当前数据中的个案创建的新个案的数量】。对由当前数据文件中的一个观测产生的新观测的数进行计数。选择此项，可以对计数变量改变默认的【变量名】和提供描述变量的【标签】。本例计数变量名称为 count，变量标签为“计数”。

⑦ 单击【下一步】按钮，打开【重组数据向导-完成】对话框，如图 2-68 所示。这是重组数据向导的最后一步，有两个选项。

- 【立即重组数据】。单击【完成】按钮后立即执行重组，将产生新的、重新构建的文件。

如果要立刻改变当前数据文件，就选择此项。

注意：如果原始数据是被加权了的，那么新数据也会是加权的，除非用作权重的变量是被重新构建的或者从新文件中去除了。

- 【将本向导生成的语句粘贴到语句窗口】。当没有准备好改变当前文件时，或者想修改语句，或者保存它以便以后再用时，选择此项。



图 2-67 【重组数据向导-第 6 步(第 7 步)】对话框

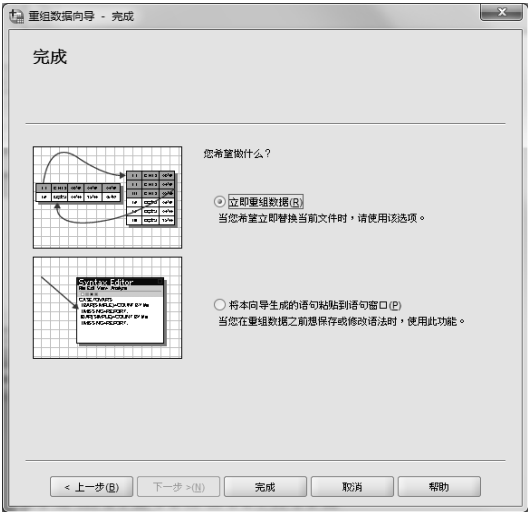


图 2-68 【重组数据向导-完成】对话框

选择①，单击【完成】按钮，程序运行转换，结果如图 2-69 所示。图 2-69(a)是索引变量值选择使用顺序值的结果；图 2-69(b)是索引变量值使用顺序值，保留变量 h、w 的结果；图 2-69(c)是第 5 步保留固定变量 h、w，索引变量值使用原始变量名的结果。运行结果保存在数据文件 data02-13-1a、data02-13-1b 和 data02-13-1c 中。

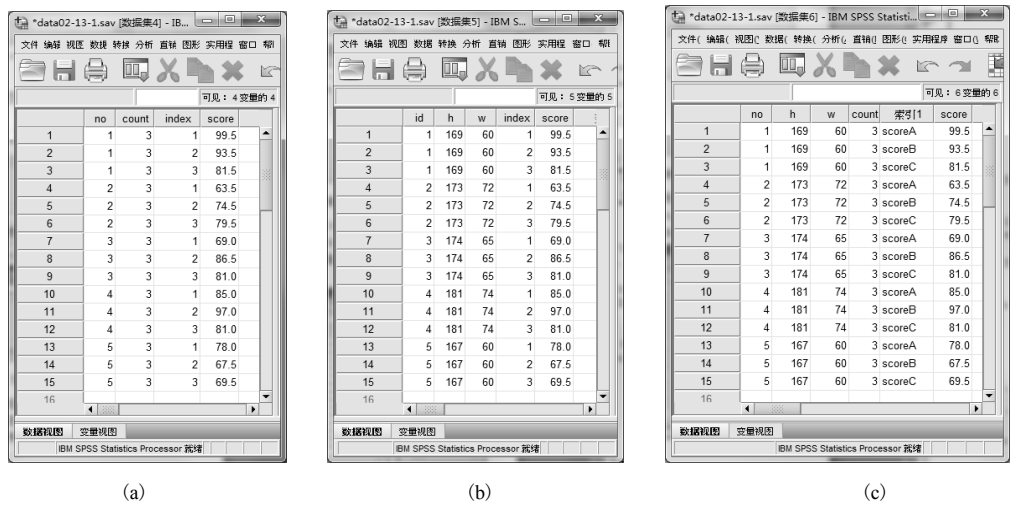


图 2-69 原始数据和不同选项生成的不同新文件

**【例 10】** 转换两个变量组，数据文件为 data02-13-2。数据包括变量：学号 no、理科成绩 scoreA1、scoreB1、scoreC1 和文科成绩 scoreA2、scoreB2、scoreC2，以及身高 h、体重 w。

转换结果如图 2-70 所示。转换步骤与上述 7 步基本相同，仅下述操作有区别：

- ① 在如图 2-63 所示的对话框中选择【多个】，而不是选择【一个】，并输入 2。
- ② 在图 2-64 中的【个案组标识】下拉菜单中选择【使用选定变量】，从当前文件的变量表中选择学号 no1 作为个案组标识变量，送入【变量】栏。

在【要转置的变量】栏的【目标变量】中输入 Lscore，将【当前文件中的变量】栏中的变量 scoreA1、scoreB1、scoreC1 送入【要转置的变量】中。就定义了第一组变量 scoreA1、scoreB1、scoreC1，新变量名为 Lscore。在【目标变量】栏下拉菜单中选择【另一项】，下面的原始变量栏自动清空。在【目标变量】栏中输入第 2 个目标变量名 Wscore。将变量 scoreA2、scoreB2、scoreC2，送入【目标变量】栏。

第 5 步选择有序数项。

其余选项不是必须改变的。原始数据如图 2-70(a) 所示，转换结果如图 2-70(b) 所示。

结果见数据文件 data02-13-2a。

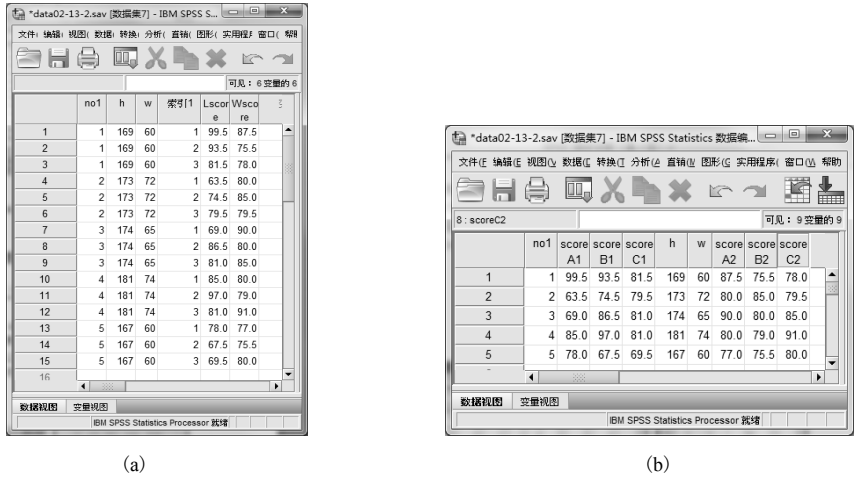


图 2-70 变换两个变量组到观测组的原始数据和新文件数据

### 【例 11】两因素不同水平的转换。

(1) 数据文件 data02-14 是三门课程 A、B、C 不同教材教学的 1、2 两个班的成绩。scoreA1、scoreB1、scoreC1 是 1 班 3 门课程的成绩, scoreA2、scoreB2、scoreC2 是 2 班 3 门课程的成绩, 如图 2-71 所示。现在进行变量组到观测的转换。要生成两个索引变量, 在分析时作为两个分类变量使用。转换结果如图 2-72 所示。

	scoreA1	scoreB1	scoreC1	scoreA2	scoreB2	scoreC2
1	99.5	93.5	81.5	87.5	75.5	78.0
2	63.5	74.5	79.5	80.0	85.0	79.5
3	69.0	86.5	81.0	90.0	80.0	85.0
4	85.0	97.0	81.0	80.0	79.0	91.0
5	78.0	67.5	69.5	77.0	75.5	80.0
6						
7						

图 2-71 两班级三课程得分原始数据

(2) 主要操作步骤如下。

- ① 在主对话框的【您希望做什么?】栏中, 仍选择第一项【将选定的变量转换为个案】, 见图 2-59。(提示: 变量要分类按序存放。)
- ② 在图 2-60 所示的【第 2 步】对话框中选择第 1 项【一个】, 一组变量将转换成一个新的因变量。
- ③ 在图 2-61 所示的【第 3 步】对话框中, 将 6 个变量 scoreA1、scoreB1、scoreC1、scoreA2、scoreB2、scoreC2 全部送入【目标变量】栏下面的【变量】栏内, 在【目标变量】栏内输入新变量名 score。在【个案组标识】栏中选择【使用个案号】。
- ④ 在如图 2-62 所示的【第 4 步】对话框中, 选择第 2 项【多个】, 并输入 2, 建立两个索引变量。也就是说, 有所选择的一组变量属于不止一个因素(条件或处理)的因变量。
- ⑤ 在如图 2-63 所示的【第 5 步】对话框中, 在表中填写变量名、变量标签和水平值。将变量名 Index1 改为 class, 在同行的标签单元中输入“班级”, 在同行的级别单元输入水平数 2; 在下一行各单元格分别输入 courses、课程、3。(提示: 此处的班级和课程顺序是对应原数据中的变量顺序的, 在原数据中变量首先按班级排序, 然后按课程排序, 所以在此处第一个索引变量是班级, 第二个索引变量是课程, 注意, 要对应, 否则结果将不是操作者所想要的。)
- ⑥ 最后一步选择第 1 项【立即重组数据】。

单击【完成】按钮, 并运行之。图 2-71 为原始数据文件, 图 2-72 为新数据文件, 重组后的数据文件详见 data02-14a。

### 【例 12】观测组到变量组结构的转换。

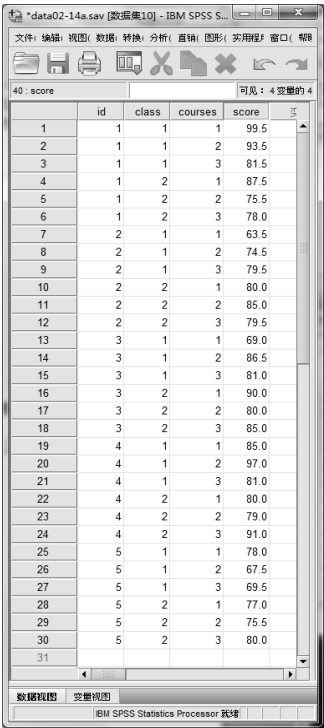
以数据文件 data02-15 为例说明转换操作。变量 time 为 2 水平的因素, 表示两个测试时间——期中和期末; courses 是 3 水平的因素, 表示 3 门课程 A、B、C。score 是分别对 5 个学生的测试成绩, 如图 2-73 所示。

这是应该进行一个因素(即 3 门课程)、两次(期中、期末)重复测量的方差分析的问题。而这个数据结构不符合分析要求, 必须转换。id 是接受测试的 5 个学生的标识变量。要按 time 变量的两个水平将分数 score 变成两个变量, 表示期中和期末对 3 门课程的测试分数, 操作如下。

(1) 在【重组数据向导】对话框(见图 2-62)中, 选择第 2 项【将选定个案重组为变量】, 即把选择的观测组变成变量组, 单击【下一步】按钮。

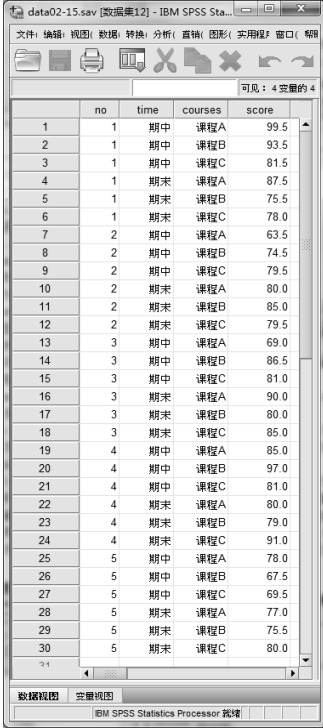
(2) 打开如图 2-74 所示的【第 2 步】对话框。将当前文件变量表中的要在新文件中作为分类变量的 courses(课程)和标识变量的 no(学生编号)送入【标识符变量】栏中; 将要按其转换的分类变量 time 送入【索引变量】栏内。单击【下一步】按钮。

(3) 打开如图 2-75 所示的【第 3 步】对话框。有两个选项，确定是否对原始数据文件按标识变量排序。第 1 项【是】，表明要求对原始数据文件按前一步指定的标识变量排序。如果没有排序或不确定是否已经排好序，应该选择此项。



	id	class	courses	score
1	1	1	1	99.5
2	1	1	2	93.5
3	1	1	3	81.5
4	1	2	1	87.5
5	1	2	2	75.5
6	1	2	3	78.0
7	2	1	1	63.5
8	2	1	2	74.5
9	2	1	3	79.5
10	2	2	1	80.0
11	2	2	2	85.0
12	2	2	3	79.5
13	3	1	1	69.0
14	3	1	2	86.5
15	3	1	3	81.0
16	3	2	1	90.0
17	3	2	2	80.0
18	3	2	3	85.0
19	4	1	1	85.0
20	4	1	2	97.0
21	4	1	3	81.0
22	4	2	1	80.0
23	4	2	2	79.0
24	4	2	3	91.0
25	5	1	1	78.0
26	5	1	2	67.5
27	5	1	3	69.5
28	5	2	1	77.0
29	5	2	2	75.5
30	5	2	3	80.0
31				

图 2-72 重组后的数据文件



	no	time	courses	score
1	1	期中	课程A	99.5
2	1	期中	课程B	93.5
3	1	期中	课程C	81.5
4	1	期末	课程A	87.5
5	1	期末	课程B	75.5
6	1	期末	课程C	78.0
7	2	期中	课程A	63.5
8	2	期中	课程B	74.5
9	2	期中	课程C	79.5
10	2	期末	课程A	80.0
11	2	期末	课程B	85.0
12	2	期末	课程C	79.5
13	3	期中	课程A	69.0
14	3	期中	课程B	86.5
15	3	期中	课程C	81.0
16	3	期末	课程A	90.0
17	3	期末	课程B	80.0
18	3	期末	课程C	85.0
19	4	期中	课程A	85.0
20	4	期中	课程B	97.0
21	4	期中	课程C	81.0
22	4	期末	课程A	80.0
23	4	期末	课程B	79.0
24	4	期末	课程C	91.0
25	5	期中	课程A	78.0
26	5	期中	课程B	67.5
27	5	期中	课程C	69.5
28	5	期末	课程A	77.0
29	5	期末	课程B	75.5
30	5	期末	课程C	80.0

图 2-73 期中、期末三门课程得分数据



图 2-74 【重组数据向导-第 2 步(共 5 步)】对话框




图 2-75 【重组数据向导-第 3 步(共 5 步)】对话框

第 2 项【否】。当原始数据文件已经按标识变量排好序了，则选择此项。系统每遇到标识变量值的一个新组合，就生成一个新行，所以，对数据文件按标识观测组的变量值排序很重要。选择后单击【下一步】按钮。打开【第 4 步】对话框，如图 2-76 所示。

- (4) 在【第 4 步】对话框中共有 3 类选项。
- ①【新变量组顺序】栏。确定新变量组的顺序，在要转换成两组以上新变量时选择才有意义。我们的例题因只生成两个新变量，故不需要选择。此类选项有两种排列方式：
- 【按初始变量排序的组合】。例如(w1、w2、w3、h1、h2、h3)。
  - 【按索引排序的组合】。例如(w1、h1、w2、h2、w3、h3)。
- ②【个案计数变量】栏，确定是否生成计数变量。选中计算当前数据中用来创新个案的个案数。在【名称】和【标签】框中分别给出变量名和变量标签。
- ③【指示符变量】栏。确定是否生成指针变量。选中【创建指示符变量】，在【根名】框中给出变量名字头。
- 重组向导可以用索引变量在新文件中生成指针变量。对索引变量的每个值生成一个新变量。指针变量指明观测的一个值出现与否。如果观测有值，则指针变量的值是 1，否则值为 0。在某些问题中，指针变量可以做频数计数用。对本例题没有用，所以不选。
- (5) 最后一步确定是立即执行，还是先生成过程语句，见图 2-68。选择第 1 项【立即重组数据】。

重组转换的结果如图 2-77 所示。详见数据文件 data2-15a。



图 2-76 【重组数据向导-第 4 步(共 5 步)】对话框

Figure 2-77 shows the data view of 'data02-15.sav' after reorganization. The table has columns 'no', 'courses', 'score.1', and 'score.2'. It displays 16 rows of data where each original case is split into two new cases based on the 'courses' variable.

	no	courses	score.1	score.2
1	1	课程A	99.5	87.5
2	1	课程B	93.5	75.5
3	1	课程C	81.5	78.0
4	2	课程A	63.5	80.0
5	2	课程B	74.5	85.0
6	2	课程C	79.5	79.5
7	3	课程A	69.0	90.0
8	3	课程B	86.5	80.0
9	3	课程C	81.0	85.0
10	4	课程A	85.0	80.0
11	4	课程B	97.0	79.0
12	4	课程C	81.0	91.0
13	5	课程A	78.0	77.0
14	5	课程B	67.5	75.5
15	5	课程C	69.5	80.0
16				

图 2-77 转换后的数据

## 2.4 观测的加权与选择

### 2.4.1 定义加权变量

在实际应用中，我们经常需要对观测进行加权处理。例如，在数据文件中如果存在一个表明相同的变量值出现频数的变量，应该定义该变量为权重变量。可以选择【数据】菜单中的【加权个案】命令，定义权重变量。至于对哪个变量的值加权，是使用权重变量计算中的问题。



1. 在选择加权变量时应该注意的事项

- (1) 权重变量中含有零、负数或缺失值的观测将被排除在分析之外；
- (2) 分数权重值有效；
- (3) 一旦定义了权重变量，那么在以后的分析中权重变量一直有效，直到取消权重变量的定义，或者定义了其他的权重变量。

2. 定义权重变量的操作



图 2-78 【加权个案】对话框

(1) 按【数据→加权个案】顺序打开【加权个案】对话框，如图 2-78 所示。

(2) 选择是否对观测进行加权处理。

①【请勿对个案加权】。是系统默认状态，表示对数据不加权，不用定义权重变量。

②【加权个案】。选择此项要求对观测加权。

(3) 选择加权变量。从左边源变量框中选择权重变量，送入【频率变量】(注：应为频数变量)框中。

(4) 单击【确定】按钮，权重变量定义完成。

2.4.2 选择参与分析的观测

如果需要部分观测参与分析，就要在分析之前进行选择，操作方法如下。

(1) 单击【数据→选择个案】打开对话框，如图 2-79 所示。

(2) 几种【选择】方法：

①【全部个案】。该选项是系统默认的，全部观测都参与分析，不做选择。

②【如果条件满足】。选择满足条件的观测。单击【如果】按钮，打开如图 2-80 所示的对话框，设置选择条件。例如只选择女性，在原变量表中选择变量 gender 送入条件编辑栏，输入“=f”，单击【确定】按钮。

③【随机个案样本】。对数据文件中的观测进行随机采样。单击【样本】按钮，打开如图 2-81 所示的二级对话框。在【样本尺寸】(应为【样本量】)栏中有两种采样方法，选择其中一种。



图 2-79 【选择个案】主对话框



图 2-80 【选择个案：if】对话框

- **【大约□所有个案的%】**。按给定的百分比近似选择。在**【大约】**后面的矩形框中输入百分比数值。
- **【精确□从第一个开始的个案□个案】**。在指定范围内随机选择给定数目的观测。在**【精确】**后面输入样本量  $n1$ ，在**【从第一个开始的个案】**后面输入一个小于或等于全部观测数的数值  $n2$ 。选择观测是从前  $n2$  个观测中选择出  $n1$  个观测。

此种方法属于重复采样，一个观测可能不只一次被选中，因此样本量与全部观测之比近似等于给定的百分比，或近似等于指定的样本量数值。

④ **【基于时间或个案全距】**。根据时间或数据范围选择，单击**【范围】**按钮打开如图 2-82 所示的对话框，输入第一个观测号和最后一个观测号，则给定范围之内的观测被选中。

⑤ **【使用筛选器变量】**。使用过滤变量，从左面的源变量栏中选择一个数值型变量作为过滤变量，过滤变量值不是 0，或不是缺失值的观测都被选中。

以上 5 种选择方法是单选项，选择一种，设置好条件参数，返回主对话框。

⑥ 在**【输出】**栏中，选择未选中的观测的处理方法。

- **【过滤掉未选定的个案】**。选择此项，未选中的观测的观测号被打上斜线，不参与分析。这是系统默认的处理方法，如图 2-83 所示。
- **【将选定个案复制到新数据集】**。选择此项，在**【数据集名称】**栏中输入新文件名。
- **【删除未选定个案】**。未选中个案的从数据文件中删除。

参数设置完成，单击**【确定】**按钮，选择完成。

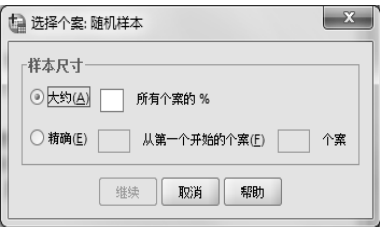


图 2-81 **【选择个案：随机样本】**对话框



图 2-82 **【选择个案：范围】**对话框

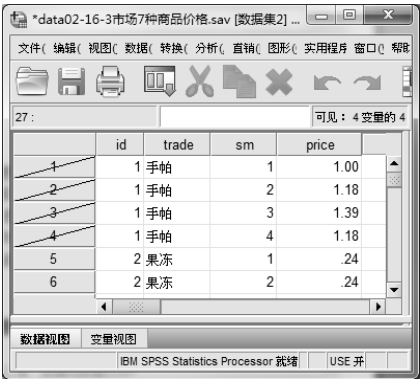


图 2-83 被滤掉的观测

## 习 题 2

1. SPSS 的变量有几种类型？
2. 变量的哪些属性会影响它们在分析中的作用？哪些属性只影响数据在窗口中的显示？哪些属性只影响输出？
3. 变量有哪几种测度方式？在分析中的作用是什么？
4. 你的工作中数据存放在什么格式的文件中？SPSS 可以直接打开这些数据文件吗？如果不能直接打开，是否能经过转换形成 SPSS 格式的数据文件？
5. 数据文件 data02-01 中，用查重功能分析受教育年限相同的职工，是否初始工资都相同？

6. 为什么要拆分数据文件？拆分的结果是什么？
7. 合并数据文件有几种情况？
8. 观测排序和排序有什么区别？什么叫结？结上观测的秩次有几种排法？体育比赛常用哪种方法排列名次？
9. 查看数据文件 `data02-03.txt`，它是否是固定格式的 ASCII 码数据文件？有列间隔吗？将其转换为 SPSS 数据文件。
10. 将数据文件 `data02-01` 按 `educ` 变量值升序排列。
11. 为什么要对变量重新编码？SPSS 有几种重新编码的过程？举例说明。
12. 什么是数据文件的重新构建？有几种重新构建的方式？
13. 超市对竞争对手的商品价格做定期调查。某天，某超市调查了 3 个竞争超市 49 种商品的售价，与本超市进行比较。从 49 种中随机抽取 7 种商品的价格，数据记录在 `data02-16` 中。因要做方差分析，要求将 4 个超市的商品价格放到一个价格变量中，另外增加变量，使数据文件能正确表达每个价格属于哪个超市、哪个商品。

# 第 3 章 输出信息的编辑

如果输出窗口中默认的常用工具不全，操作不方便。最好使用【视图】菜单中的【工具栏】功能将常用工具按钮显示在工具栏中，见第 1 章 1.2.6 节内容。例如，在默认工具栏中加入【剪切】、【复制】、【粘贴】、【删除】等常用图标按钮。

SPSS 使用与 Windows 系统相同的基本编辑功能和图标按钮，查找与替换的操作与 Windows 系统的同类功能一致，本章只介绍输出窗口中的一些特殊的常用编辑方法。

## 3.1 输出窗口中的文本浏览与编辑

系统中的操作与过程运行的结果显示在输出窗口中。输出窗口的导航系统是比较特殊的输出窗口信息浏览器，它不但为窗口中的内容查找、浏览提供了工具，同时也为窗口中的内容编辑提供了方便。

### 3.1.1 利用导航器浏览输出信息

图 3-1 所示是输出窗口。左半部分是导航器，右半部分是输出信息区。导航器实际上是可折叠的输出信息树形结构图。对导航器中的每一项都可以使用鼠标左键单击进行选择，双击进行打开与隐藏的操作。

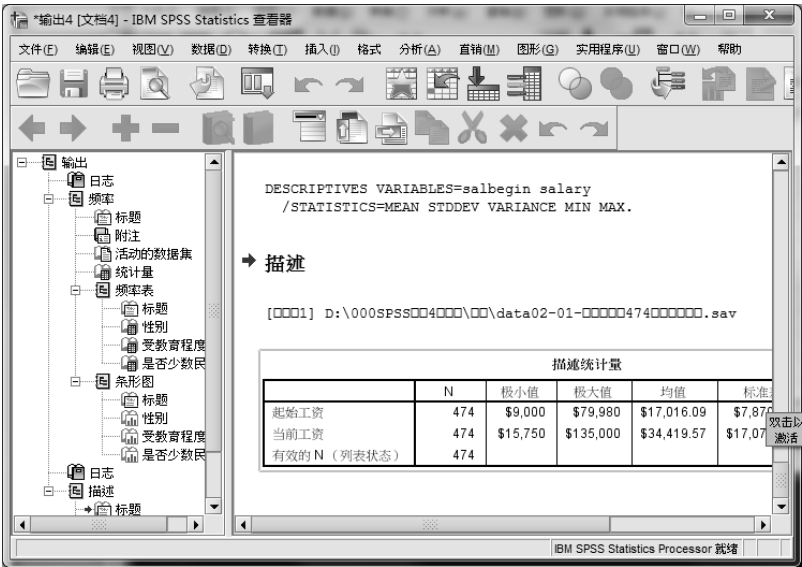


图 3-1 输出窗口

### 1. 认识导航器

导航器中有【输出】总项，这是最高一级的输出项；以过程语句命名的过程项，如图 3-1

中的频率、描述过程项是第二级；第一、二级输出项前显示的是带有结构图的书形图标，而且有折叠图标(加减号)。每个过程项都可能包括几种结构项，即第三级结构项。这些结构项是否显示在导航器中，取决于系统参数设置，参见 1.3 节有关内容。可能显示的结构项有：日志项、标题项、附注说明项、活动数据集(当前的工作数据集)、统计量、表格(如图 3-1 中的频率表统计量表)、警告项、统计图项(如图 3-1 中的条形图等)和文本输出项。

2. 在导航器中选择输出项

单击导航器结构图中的某一项，与该项相应的输出信息显示在右侧窗口的可见部位，且外轮廓加了黑实线框。用这种方法可以将需要浏览的部分调入窗口中的可见信息区。

3. 在导航器中关闭/打开输出项

为突出浏览重点，可隐藏部分输出线，需要浏览时再显示出来，操作如下。

(1) 导航窗口中的一级控制项是输出项，单击前面的加(或减)号图标，所有输出内容全部显示(或隐藏)，如图 3-2(a) 所示。

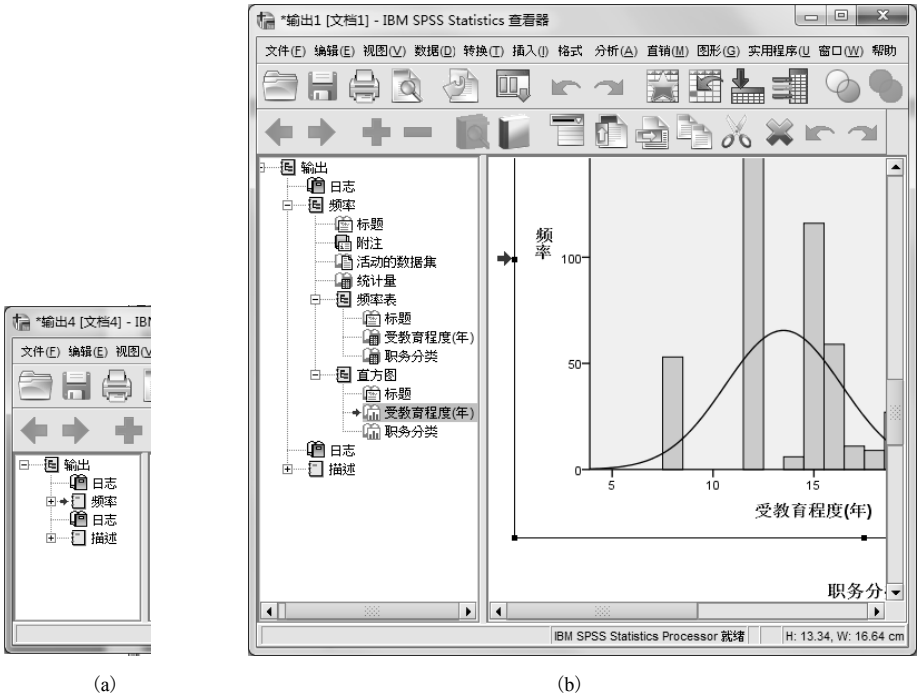


图 3-2 输出项的打开与隐藏

(2) 隐藏/显示各级内容。导航窗口中第二级是过程控制，每一项是一个过程输出，单击过程项前面的加(减)号图标，可以显示(或隐藏)过程项中的内容，图 3-2(a)、图 3-2(b)所示分别是二级项隐藏和打开的状态，描述是被隐藏了二级结构项的过程项，而频率过程项是被打开的。二级项打开，各项前有加减号图标的是三级项。例如，频率二级项下面的频率表就是三级项。三级项的下一级各项前有书形图标或输出类型图标，单击该图标，左侧产生红色箭头，其内容在右侧窗口中显示，例如图 3-2(b)所示的受教育程度条形图。

(3) 在右面信息窗口中显示某项输出的方法是，在导航器中单击各级项前的加号图标，展

开，直到显示出带有书形图标的各项。单击使某项前面出现红色箭头，书形图标打开，相应内容在右侧窗口可见。在对应的信息(一个标题、一个表格、一个统计图或一段完整的文本)左边也显示出红色箭头，外边显示黑色框线。

如果第 3 级不止一个输出项，例如图 3-2(b)中的条形图有 4 个，鼠标单击一项，右侧窗口显示相应的输出项。

3.1.2 编辑导航器中的输出项

(1) 选择操作对象

使用【编辑】菜单的【选择】子菜单中的【选择】功能对操作对象进行分类选择，如图 3-3 所示。子菜单中的各项选择功能如下。

- 【最后的输出】。指最后一次执行 SPSS 过程的全部输出。
- 【所有标题】。指只选择各次执行 SPSS 过程的所有标题。例如，需要同时调整所有标题的字体、字号时，可以选择此项。
- 【所有页面标题】。例如，需要改变所有页面标题的字体字号时，可以选择此项。
- 【所有枢轴表】。即选择各次执行 SPSS 过程的所有输出表格。例如，需要同时改变所有表格的格式或只复制所有表格时，可以选择此项。
- 【所有图表】。例如，需要同时改变所有统计图表的元素属性或只复制所有图表时，选择此项。
- 【所有文本输出】。例如，需要修改所有输出文本的字体、字号等时，可以选择此项。

后面各项所指的选择内容可以做类似理解。

- 【所有警告】信息。
- 【所有注释】。即所有说明信息。
- 【所有日志】。
- 【所有其他对象】。也就是上述各项未包括的信息。
- 【所有树】(形图)。
- 【所有模型】。

(2) 使用鼠标键选择操作对象

- ① 可以用鼠标左键在导航器中单击一个结构项，选择一个操作对象，使之彩底显示。
- ② 按住 Ctrl 键的同时，鼠标左键单击要选择的对象，可选择位置不连续的多个对象。
- ③ 按住 Shift 键的同时，用鼠标左键分别单击两个不相邻的结构项，可以选择这两个结构项之间的各项。

(3) 对选中的结构项及其内容可以进行删除、剪切到剪贴板、复制到剪贴板和粘贴到另一位置的操作。

(4) 移动输出项显示位置的另一种方法是，用鼠标拖动到目标位置。用这种方法可以将有用信息组织到一起，建议只在一个过程项内做这种操作，以免出现混乱。



图 3-3 【编辑】菜单和【选择】子菜单

### 3.2 输出表格中信息的编辑

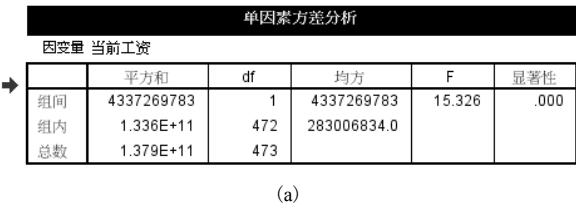
#### 3.2.1 表格编辑工具与常用编辑方法

##### 1. 选择操作对象

要编辑某个表格，必须先选择它。双击要编辑的表格即选择了这个表格，被选中的表格的标题是反向显示的，左侧显示红色箭头，右侧、下方均显示虚线，如图 3-4(a) 所示。

在表格中具体选择表格元素，使用下面的方法：

- (1) 选择一个单元格，只要鼠标左键单击这个单元格即可。
- (2) 选择两个以上单元格，需要按住 Ctrl 键，用鼠标单击需要选择的各单元格。



(a)

	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
否	370	\$36,023.31	\$18,044.096	\$938.068	\$34,178.68	\$37,867.94	\$15,750	\$135,000
是	104	\$28,713.94	\$11,421.638	\$1,119.984	\$26,492.72	\$30,935.17	\$16,350	\$100,000
总数	474	\$34,419.57	\$17,075.661	\$784.311	\$32,878.40	\$35,960.73	\$15,750	\$135,000

(b)

图 3-4 选择表格中的操作对象

(3) 选择一行或一列，只要用鼠标拖动经过所需要选择的行或列，或者按住 Ctrl 键单击并拖动相邻的两个以上单元格，就选中了这两个单元格所在的行或列。

选择不只一行或一列，可以按住鼠标拖动，所经过的行(或列)均反向显示。

(4) 在有的表格中，选择一行(或一列)会导致把与之相关的行(或列)也同时选择，如图 3-5(c) 所示。

##### 2. 表格编辑工具

选择了要编辑的表格，单击【视图】菜单，在二级菜单项中单击【工具栏】，会显示表格编辑工具栏，如图 3-5 所示。选择了表格中的编辑对象，工具栏中可用的工具会亮。

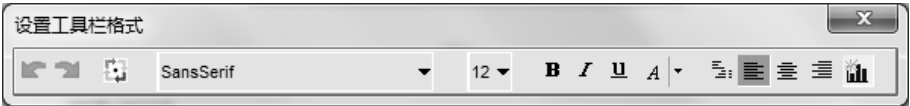


图 3-5 表格编辑工具栏

表格编辑工具栏工具及功能除表格翻转控制功能外，其他功能(如撤销与恢复操作、单元格中字体字号的设置、对齐方式的设置)都与 Windows 中相应功能的操作方法相同。

3. 标题与文字编辑

双击一个标题或表格中的文字，如图 3-6 所示，被编辑的标题变成蓝底白字(默认)，即可使用表格工具进行编辑，包括修改文字内容、改变字体、字号、对齐方式等。表 3-6(a)所示为选择了要进行比较的表格标题，表 3-6(b)所示为表格标题变为 18 号华文新魏体，并加粗、倾斜的结果，且选择了要进一步编辑的表头文字。

单因素方差分析						单因素方差分析					
因变量 当前工资						因变量 当前工资					
	平方和	df	均方	F	显著性		平方和	df	均方	F	显著性
组间	4337269783	1	4337269783	15.326	.000	组间	4337269783	1	4337269783	15.326	.000
组内	1.336E+11	472	283006834.0			组内	1.336E+11	472	283006834.0		
总数	1.379E+11	473				总数	1.379E+11	473			

表 3-6 表格中的文字编辑

4. 修改单元格中的内容

修改表格单元格中的任何内容均可以仿照修改表格标题的操作方法，但最好不要修改单元格中的数据。除了应该实事求是外，还因为修改一个数据会影响其他数据的正确性，这是 SPSS 的输出与 Excel 工作表的不同之处。例如在图 3-7 中，将变量当前工资描述统计量表中的 N 观测总数改变后，与之有关的统计量(如均值)数值不会自动随之改变，因此整个表格中的其他值就全部错了。

因变量 当前工资								
	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
否	370.0000000	\$36,023.31	\$18,044.096	\$938.068	\$34,178.68	\$37,867.94	\$15,750	\$135,000
是	104	\$28,713.94	\$11,421.638	\$1,119.984	\$26,492.72	\$30,935.17	\$16,350	\$100,000
总数	474	\$34,419.57	\$17,075.661	\$784.311	\$32,878.40	\$35,960.73	\$15,750	\$135,000

图 3-7 修改表格中的数据会造成错误

5. 隐藏或显示表格的行与列

- (1) 双击选定的表格，选择要隐藏的行或列，图 3-8(a)所示即选择了一行。
- (2) 单击【视图】菜单，选择【隐藏】命令，选定的表格或栏目被隐藏起来，如图 3-8(b)所示。
- (3) 如果要恢复显示被隐藏的行或列，必须先选择邻近的未隐藏的行或列，再选择【视图】菜单中的【显示】所有类别命令。

选择表格或某一部分后，可以选择的【视图】菜单如下：

- ① 【隐藏】。隐藏所选择的表格元素。
- ② 【显示位数标签】(SPSS 汉化为【显示维数标签】)。
- ③ 【隐藏所有类别标签】。
- ④ 【显示所有类别】。
- ⑤ 【显示所有注脚】。
- ⑥ 【全部显示】。



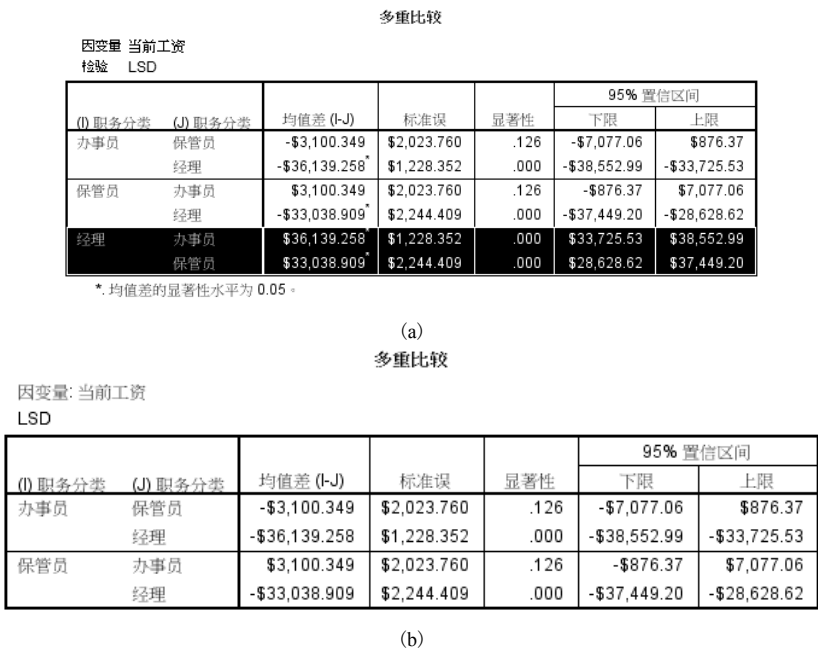


图 3-8 表格内容的选择与隐藏

6. 改变表格列宽度

表格的显示常常因列宽不够而将数字显示成一系列星号，这就需要调整表格的列宽。

(1) 手动调整

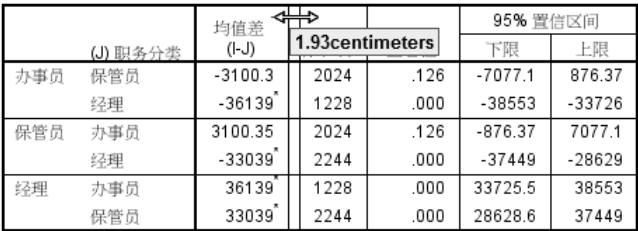
将鼠标光标置于要调整的表格竖线上，此时光标变为水平的双箭头线，同时待调整的竖线加粗。按住鼠标左键，拖动鼠标调整表格列宽，直到列宽合适，或未显示的数值显示出来为止，松开鼠标左键，如图 3-9 所示。

应该注意，调整列宽时显示的列宽数值，可作为调整参考。列宽的调整会影响列中数据显示的有效数字位数。当调整列宽过小时，会显示出“隐藏”字样，如果此时松开鼠标左键，列宽过小的栏目会被隐藏；如果发现因宽度调整而隐藏了一列不应隐藏的数据，可以使用【编辑】菜单中的【撤销】命令将其撤销。

(2) 菜单命令调整

双击选择表格，按【格式→设置数据单元格宽度】顺序单击菜单项，打开如图 3-10 所示的对话框。在【所有数据单元格的宽度】后面的编辑区中输入宽度值，也可以单击向上、向下箭头按钮增加或减少宽度数值。单击【确定】按钮确认。

这样设置的宽度产生的效果是除最左面一列外，其他各列等宽。



\*. 均值差的显著性水平为 0.05。

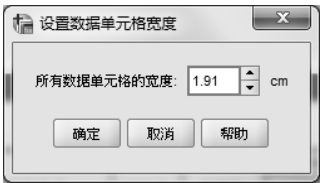


图 3-9 手动调整表格的列宽

图 3-10 【设置数据单元格宽度】对话框

选择了整个表格后如果选择【格式】菜单中的【自动调整】，则各列宽度按数据的宽度自动调整，而不必每列分别调整。

3.2.2 表格的转置与行、列、层的处理

表格是运行分析过程自动产生的。表格形式不一定能满足编写报告的要求，例如，行、列的安排使得表格过长或过宽，都会在一定程度上影响对数据的观察和分析。可以用下面的方法进行转换。

1. 使用菜单对表格进行行、列互换(转置)

- (1) 双击选定的表格，使之显示外阴影框。此时输出窗口中的主菜单发生改变。
- (2) 按【枢轴→转置行和列】顺序单击菜单项。图 3-11 (a)所示为对一个频数分布交叉表进行转置，转置后的结果如图 3-11 (b)所示。

性别\*受教育程度(年) 交叉制表

		性别		合计
		女	男	
受教育程度(年)	8	30	23	53
	12	128	62	190
	14	0	6	6
	15	33	83	116
	16	24	35	59
	17	1	10	11
	18	0	9	9
	19	0	27	27
	20	0	2	2
	21	0	1	1
	合计	216	258	474

性别\*受教育程度(年) 交叉制表


		受教育程度(年)										合计
		8	12	14	15	16	17	18	19	20	21	
性别	女	30	128	0	33	24	1	0	0	0	0	216
	男	23	62	6	83	35	10	9	27	2	1	258
合计		53	190	6	116	59	11	9	27	2	1	474

(a) (b)

图 3-11 转置前后的表格


2. 使用表格转置盘进行行、列、层之间的位置转换

使用菜单对表格只能进行行、列之间的转置，如果表格还有第三维层，在行、列、层之间的转换则应该使用表格托盘。

在输出信息区，双击要进行编辑的表格，再按【枢轴→透视托盘】顺序单击菜单项；或者在出现的表格编辑工具条中单击表格转置工具盘图标按钮，都会打开【透视托盘】对话框。如图 3-12 (b)所示，托盘标题栏标有“透视托盘”，也称转置盘。

3. 利用转置盘变换层、行、列的位置

使用鼠标拖曳层标、行标、列标中的任意一个变量到另一个位置，可以改变层、行、列的相互关系，使层、行、列上的数据转换位置，使表格满足显示要求。

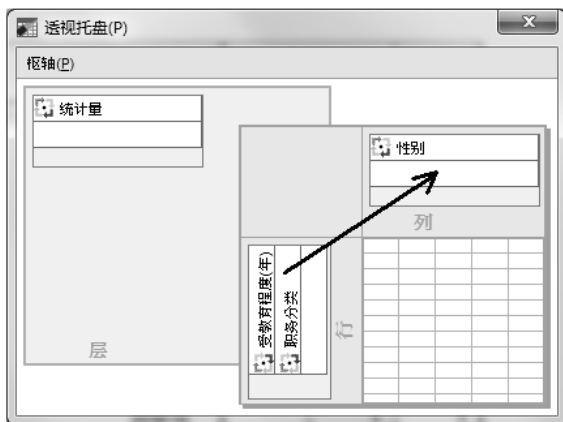
图 3-12 (a)是一个很长的表格。图 3-12 (b)是双击表格后打开的【透视托盘】对话框。托盘有 4 个图标，分别代表层、行、列上的变量或统计量。图 3-12 (b)托盘中的表格左边标有“行”处是行标，行标上的变量是“受教育程度”和“职务分类”；在转置盘右上方标有“列”处的是列标，列标上的变量是“性别”。如果信息区中选择的是二维表格，则只有行标和列标。此为三位表格，层标上是“统计量”。

鼠标左键点行变量“职务分类”，按住鼠标左键，将该变量拖至列标第 2 行上，如图中箭头所示。拖曳结果如图 3-12(c)所示，松开鼠标键就可以看到变化了的表格。观察变化后的表格是否符合要求，如果不满意还可以将变量拖回行标。拖曳后表格如图 3-12(d)所示，显然，横向长的表格更易于在文章中排版。

受教育程度(年)\* 性别\* 职务分类 交叉制表

计数		性别			
职务分类		女	男	合计	
办事员	受教育程度(年)	8	30	10	40
		12	128	48	176
		14	0	6	6
		15	33	78	111
		16	14	10	24
		17	1	2	3
		18	0	2	2
		19	0	1	1
合计		206	157	363	
保管员	受教育程度(年)	8		13	13
		12		13	13
		15		1	1
	合计			27	27
经理	受教育程度(年)	12	0	1	1
		15	0	4	4
		16	10	25	35
		17	0	8	8
		18	0	7	7
		19	0	26	26
		20	0	2	2
		21	0	1	1
	合计		10	74	84
合计	受教育程度(年)	8	30	23	53
		12	128	62	190
		14	0	6	6
		15	33	83	116
		16	24	35	59
		17	1	10	11
		18	0	9	9
		19	0	27	27
		20	0	2	2
		21	0	1	1
	合计		216	258	474

(a)



(b)



(c)

受教育程度(年)\* 性别\* 职务分类 交叉制表

计数		性别											
		女			男					合计			
		职务分类			职务分类				职务分类				
		办事员	经理	合计	办事员	保管员	经理	合计	办事员	保管员	经理	合计	
受教育程度(年)	8	30		30	10	13		23	40	13		53	
	12	128	0	128	48	13	1	62	176	13	1	190	
	14	0		0	6			6	6			6	
	15	33	0	33	78	1	4	83	111	1	4	116	
	16	14	10	24	10		25	35	24		35	59	
	17	1	0	1	2		8	10	3		8	11	
	18	0	0	0	2		7	9	2		7	9	
	19	0	0	0	1		26	27	1		26	27	
	20		0	0			2	2			2	2	
	21		0	0			1	1			1	1	
合计		206	10	216	157	27	74	258	363	27	84	474	

(d)

图 3-12 用表格托盘对带层的表格进行转置

4. 层变量位置的变换

为了便于观察，层变量一般放在表格左上角，如果层变量有两个以上类别，会形成下拉列表形式。

图 3-13 (a) 相对于图 3-12 (b) 的变化为，将“职务分类”变量拖至层标上，则图 3-12 (a) 表格转换成图 3-13 (b) 的形式。

如图 3-13 (b) 所示，作为报告的一部分，下拉列表不能起作用，但是可以把一个大表分成 3 个较小的表。在行列菜单中每选择一个职务就显示出一个小表格。相反转换，把在行或列上的分类变量拖曳到层上生成层，对观察输出结果也有很大好处。



(a)

受教育程度(年)\* 性别\* 职务分类 交叉制表

统计量	计数	性别		
		女	男	合计
职务分类	办事员			
	保管员			
	经理			
受教育	合计			
	8	30	10	40
	12	128	48	176
	14	0	6	6
	15	33	78	111
	16	14	10	24
	17	1	2	3
	18	0	2	2
	19	0	1	1
合计		206	157	363

(b)

图 3-13 “职务分类”变量转至层标上及转换后的表格

3.2.3 表格外观的设置与编辑

1. 表格样式设置

SPSS 为读者预设了一些格式的表格，每个格式的表格都有各自的特点，读者可以根据需要选择这些表格的样式。

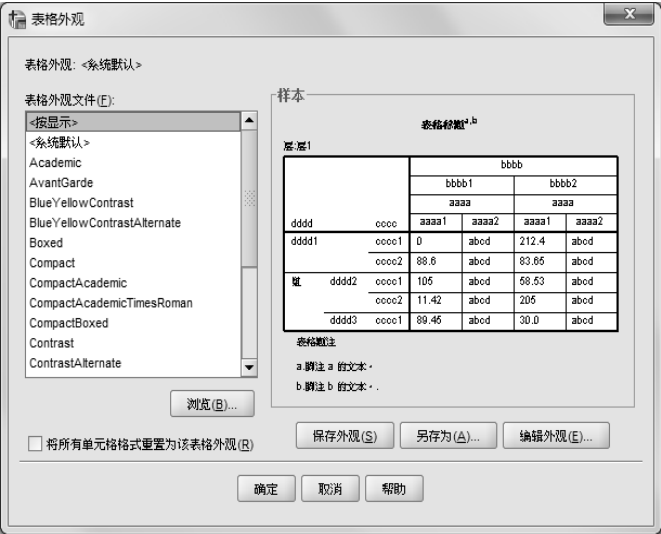


图 3-14 【表格外观】对话框

一般外观的特征设置如下：

① 双击表格使其进入编辑状态。

② 按【格式→表格外观】顺序单击菜单项，打开相应的对话框，如图 3-14 所示。

③ 在【表格外观文件】栏中选择表格样式文件，在【样本】栏中观察所选定的表格样式文件代表的表格样式。

④ 如果要选择保存在文件中的其他表格样式，则单击【浏览】按钮，打开【文件打开】对话框，选择文件。SPSS 表格样式文件是 tlo 格式，文件扩展名为 tlo。

⑤ 将所有单元格格式重置为该表格外观。在编辑过的表格中选择这个选项，将废除原来的编辑结果，将表格中所有编辑过的单元格重新设置成这里选择的表格样式。

⑥ 单击【保存外观】按钮打开【保存】对话框。将当前选择的表格样式保存为当前选择的表格样式文件，以备需要时使用。

⑦ 单击【另存为】按钮打开【另存为】对话框，将当前选择的表格样式保存到指定路径下的指定文件中。

⑧ 单击【编辑外观】按钮，显示【表格属性】对话框。在该对话框中可以按需要对表格属性对话框中选择的表格样式进行修改和编辑。

⑨ 单击【确定】按钮，所选择的表格变成在对话框中选择的样式。

2. 表格样式的编辑

要对表格样式进行编辑，可以先使用【表格外观】对话框选择一种基本样式，然后对所选择的样式进行修改。进入【表格属性】对话框的途径有两个：

- 从【表格外观】对话框中选择一种表格样式后，单击【编辑外观】按钮打开【表格属性】对话框，如图 3-15 所示。
- 按【格式→表格属性】顺序单击菜单项，打开【表格属性】对话框。

对话框中有【常规】、【脚注】、【单元格格式】、【边框】、【打印】5 个功能选项卡，可从这 5 个方面修饰表格。



图 3-15 【表格属性】对话框

(1) 常规特性设置

① 在对话框【常规】选项卡的【常规】栏中选择【隐藏空行和空列】。表格中的空行或空列将不显示。

② 在【行维数标签】栏内设置作为维度的变量，例如方差分析中的因子变量的变量名和变量标签显示的位置，有两种方式：【内角】与【嵌套】，如图 3-16 所示。

- 【内角】。变量名显示在表格左上角单元格中。变量标签显示在变量名旁边，一般显示在变量名右边，如图 3-16(a) 所示。如果在系统参数设置中设置了输出只显示变量标签，则不会显示变量名。

- 【嵌套】。以嵌套方式显示，如图 3-16(b)所示。

				gender 性别	
				女	男
jobcat 职务分类	educ 受教育程度(年)	8		30	10
办事员		12		128	48
		14		0	6
		15		33	78
		16		14	10
		17		1	2
		18		0	2
		19		0	1
合计					
保管员	educ 受教育程度(年)	8			

(a)

				gender 性别	
				女	男
jobcat 职务分类	办事员	educ 受教育程度(年)	8	30	10
		12		128	48
		14		0	6
		15		33	78
		16		14	10
		17		1	2
		18		0	2
		19		0	1
合计				206	
保管员	educ 受教育程度(年)	8			

(b)

图 3-16 行维度标签显示位置

- ③ 在【列宽度】栏中以像素点为单位设置列、行的极限宽度。
- 【列标签最小宽度】。设置值可以在后面的矩形框中输入，或单击上下箭头按钮改变设置值。
  - 【列标签最大宽度】。框中设置最大列宽度，此值必须大于最小列宽度。
  - 【行标签最小宽度】。框中设置最小行宽度，此值必须小于最大行宽度。
  - 【行标签最大宽度】。框中设置最大行宽度，此值必须大于最小行宽度。
- 设置完成后，单击【确定】按钮，对所选择的表格即刻产生效果。

(2) 脚注设置

在输出的表格中常常需要添加脚注。在【表格属性】对话框中的【脚注】选项卡中进行设置，如图 3-17 所示。

- ① 在【编号格式】栏，设置脚注方式。在【样本】栏内看设置的实际效果。
- 【字母顺序】。使用字母作为脚注标记，按顺序排列，第一个脚注用 a，第二个脚注用 b...
  - 【数值】。使用数字作为脚注标记，顺序为 1、2、3、4...



图 3-17 【脚注】选项卡

② 在【标记符位置】栏，设置脚注位置。

- 【上标】。脚注标记为上标，显示在被标记对象的右上角。
- 【下标】。脚注标记为下标，显示在被标记对象的右下角。

图 3-17(a)所示为【表格属性】对话框中的【脚注】选项卡。图 3-17(b)所示为 4 种脚注方式样例。第一个脚注设置的是用字母顺序上标；中间的样例是数值上标；最下面一个是字母下标。

单击【应用】按钮，再单击【确定】按钮，设定的脚注即刻对所选择表格中的脚注生效。

(3) 单元格格式的设置

单击【单元格格式】选项卡，在该选项卡中设置单元格格式，如图 3-18 所示。

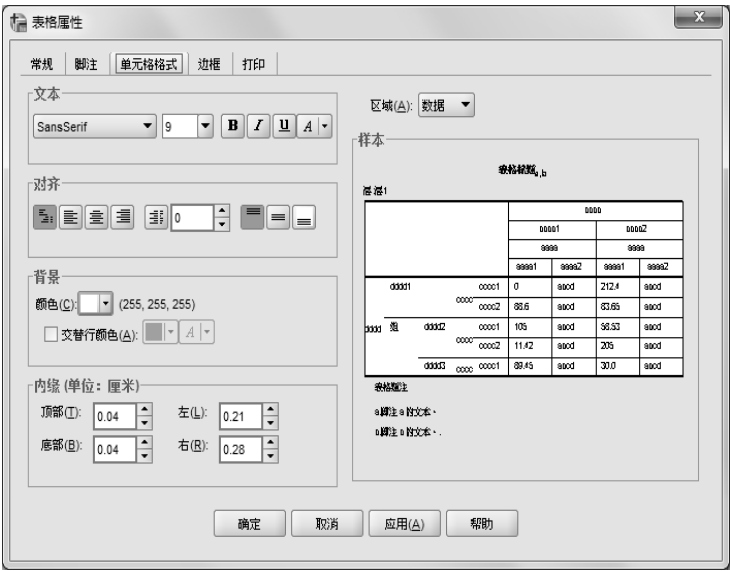


图 3-18 【单元格格式】选项卡

① 【区域】框中选择要编辑哪个区域的单元格格式：【标题】、【图层】、【(左上)角标签】、【行标签】、【列标签】、【数据】、【题注和脚注】。

可以通过【文本】栏设置字体颜色，在【样本】区域中可看到以上各项代表的表格区域。

② 在【文本】栏设置字体、字号、加粗、倾斜、加下画线和改变字体颜色。

③ 【对齐】栏，设置表格中指定元素的对齐方式，按钮自左至右分别为：

- 混合对齐。数字、日期右对齐，所选择的其他元素在单元格中左对齐。
- 左对齐。所选区域中的文字、数字对齐到单元格的左边界。
- 居中。所选区域中的内容居中对齐。
- 右对齐。所选区域中的文字、数字对齐到单元格的右边界。
- 小数对齐。所选区域中的小数点距右边界的距离为指定的距离。
- 指定单元格中数字的小数点与右边界的距离，单位是点、英寸或厘米，这个单位在【选项】对话框的【常规】选项卡中指定；见 1.3 节。
- 顶端。所选区域单元格中的内容对齐到上边界。
- 居中。所选区域单元格中的内容垂直居中。
- 底端。所选区域单元格中的内容对齐到下边界。

④【背景】栏中设置背景颜色。单击【颜色】框中的【向下箭头】按钮，显示调色板，选择背景颜色。

⑤ 在【内缘(单位：厘米)】栏中设置单元格内容与顶部、底部、左(边界)和右(边界)的距离。

(4) 设置边框格式

单击【边框】选项卡，如图 3-19 所示，设置表格边框格式。表格边框指表格各位置上的表格线。

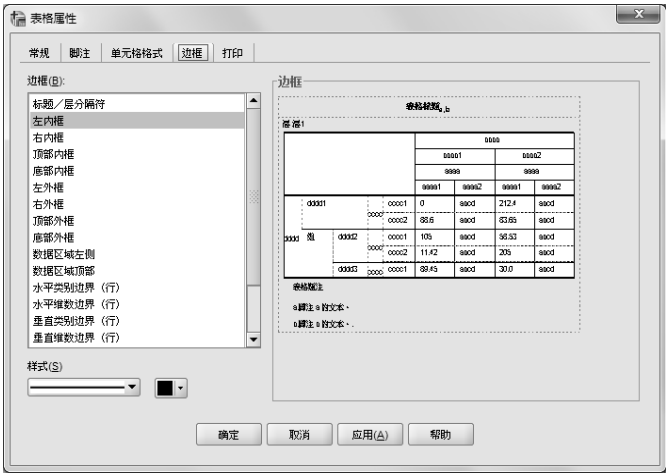


图 3-19 【边框】选项卡

选项卡左边【边框】栏中列出的是各边框线的名称，在该栏中选择一种表格线，在【样式】栏中就显示当前该位置上的表格线线型，单击右侧的向下箭头按钮，在下拉列表中选择一种线型，在栏下面左侧的下拉列表中选择线型，在右边的下拉列表中选择线的颜色。在设置了不同的线型或颜色后，可以在右边的(预览)【边框】栏中看出选项指的是哪些边框线，以及所设置的效果。单击【应用】按钮，可以在输出窗口中看到所选择表格的设置效果。

(5) 设置打印参数

单击【打印】选项卡，如图 3-20 所示，设置有关打印的参数，其选项及其含义如下：



图 3-20 【打印】选项卡



①【打印所有层】上的表格。选择此项，激活【在单独页上打印各层】复选项，如果选择该复选项，则各层表格打印在一页上。

②【调整宽表格比例以适合页面】。压缩一个过宽的表格，保持表格的纵横比，以便在打印时适应在页面设置中设置的页宽。

③【调整长表格比例以适合页面】。压缩一个过长的表格，保持表格的纵横比，以便在打印时适应在页面设置中设置的页长。

④【窗口/孤行】。如果一个表格对所设置的页来说太长或太宽时，该设置可以规定一个打印区中包含的最小行数和列数。

⑤【连续文本】。后面的编辑区中输入一个标志性文字，默认的是“Cont.”。当要打印的表格对设置的页来说太长，需要打印在两页上时，将连续文本后面编辑区中的文字打印在两页接续之处。

⑥【连续文本位置】栏，设置在⑤中定义的接续文字显示(打印)的位置。表示接续的文字只在打印时出现，也可以使用打印预览观察到。

- 【在表格底部(题注末端)】。表示接续的文字显示在表格底部。如果表格已经有了标题，则接续文字加在标题后面。
- 【在表格顶部(题注末端)】。表示接续的文字显示在表格顶部。如果表格已经有了标题，则接续文字加在标题后面。

### 3.2.4 输出信息的复制与打印

如果撰写论文需要的分析结果数据在输出表格中，可将文字、表格复制到用 Word 撰写的论文中，可以使用选择、复制、粘贴的方法，但要注意以下两点：

(1) 直接将表格粘贴到 Word 文档中，其结果仍是 Word 表格，可以使用 Word 表格功能进行编辑和调整。



图 3-21 【选择性复制】对话框

(2) 可以在所选表格中单击鼠标右键，在右键菜单中选【选择性复制】，打开【选择性复制】对话框，如图 3-21 所示。在对话框中将所有项均选中，并选择【保存为默认】项。

在 Word 中可以在粘贴表格时，在插入点处单击鼠标右键，选择右键菜单中的【选择性粘贴】。这时就可以根据需要进行任何形式的粘贴了。

如果选中表格后，单击右键菜单中的【复制】项，再粘贴到 Word 文档中，则这个表格也是图片格式，可以使用 Word 中图片工具栏中的各种工具对表格进行编辑。

**注意：**如果表格太宽，可以在复制之前先调整表格宽度，或隐藏不必要的数据列。

打印的参数设置与操作可以参考 Windows 的打印设置与操作。

## 习 题 3

1. 导航器的作用是什么？
2. 输出表格的数据能任意改变吗？为什么？
3. 怎样组织输出内容？
4. 表格太长，一页的宽度不能显示全部内容怎么办？

# 第 4 章 随机变量与分布函数的应用

## 4.1 随机变量与分布函数

### 4.1.1 随机变量及其概率分布

#### 1. 随机变量

按照机会或概率取值的变量称为随机变量。

研究随机变量要对其取不同值的随机事件进行分析，比如，掷骰子时，用点数表示 6 个随机事件；合格产品用 1 表示，不合格产品用 0 表示；对某产品喜欢用 1 表示，不喜欢用 0 表示；喜欢程度：很喜欢、喜欢、无所谓、不喜欢、很不喜欢分别用 1、2、3、4、5 表示。

随机变量根据其取值的类型分为离散型随机变量和连续型随机变量。

随机变量取有穷多值或可列无穷多值的称为离散型随机变量。例如，连续射击直至命中时的射击次数，其可能取的值为 1, 2, 3, ...。

可以取某区间中或某些区间中任何值的随机变量称为连续型随机变量。例如，1000 个成人样本中的身高可以取 1.4~2.0 m 之间的任何值，体重可能取值在 30~100kg 之间等。

#### 2. 离散型随机变量的概率分布

离散型随机变量的取值是有限的或可列无限的，如果知道每个可能取值的概率，就可以用表格、图形(如表示相对频数的柱形图)或公式、表格表达概率分布的状况。

离散型随机变量的概率分布表示为：

设  $x$  所有可能的不同取值为  $x_i, i=1, 2, \dots, n$ ，或可列无限的  $i=1, 2, \dots$

$P(x_i) = p_i, i=1, 2, \dots, n$ ，对可列无限的  $i=1, 2, \dots$

离散型随机变量的重要性质：

$$p_i \geq 0, \quad \sum_{i=1}^n p_i = 1$$

常见的离散型随机变量的概率分布有以下 3 种。

##### (1) 两点分布

两点分布又称伯努利分布，是二项分布的特例。所谓的伯努利试验，是指独立地实行只有两种可能结果，且出现某种结果概率不变的试验。重复实验只有两种互斥的事件，即事件的发生与不发生，那么这两种事件的分布服从两点分布。也就是说，随机变量只能取两个值，事件发生则取值 1，不发生则取值 0。如抛掷硬币的正面与反面；市场调查中对一件商品的态度：购买与不购买等。两点分布的概率分布函数表示为：

$$P(X=x) = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases}$$

与之有关的函数为:

- PDF.BERNOULLI(*quant*, *prob*) 数值型函数, 函数值等于分布参数为 *prob* 的伯努利分布在 *quant* 处的概率值。
- CDF.BERNOULLI(*quant*, *prob*) 数值型函数, 给出符合概率为 *prob* 的二项分布的随机变量值小于或等于 *quant* 的累积概率值。
- RV.BERNOULLI(*prob*) 数值型函数, 函数值为一个来自伯努利分布且具有指定概率参数 *prob* 的随机数。

## (2) 二项分布

满足下列条件的分布为二项分布:

- ① 从总体中抽取  $n$  个单元组成样本(即重复  $n$  次实验)。
- ② 各次实验相互独立, 每次实验只能有两种互斥的结果, 即某事件  $A$  发生与不发生。
- ③ 每次实验, 事件  $A$  发生的概率为  $\pi$ , 记做  $P(A) = \pi$ ; 不发生的概率为  $1 - \pi$ 。在  $n$  次实验中事件  $A$  发生次数的概率的分布为二项分布, 如果用  $X$  表示事件  $A$  发生次数的随机变量, 则该概率的表达式为:

$$P(x) = P(X = x) = C_n^x \pi^x (1 - \pi)^{n-x} \quad X = 1, 2, 3, \dots, n$$

SPSS 的二项分布概率函数为 PDF.BINOM 函数。

- PDF.BINOM(*quant*, *n*, *prob*) 数值型函数, 每次实验成功的概率是 *prob* 时, 函数值为  $n$  次实验中的成功次数等于 *quant* 的概率。当  $n = 1$  时同 CDF.BERNOULLI 函数。
- CDF.BINOM(*quant*, *n*, *prob*) 数值型函数, 当每次实验成功的概率是 *prob* 时, 函数值是一个  $n$  次实验中成功次数小于或等于 *quant* 的二项分布累积概率值。当  $n = 1$  时同 CDF.BERNOULLI 函数。
- RV.BINOM(*n*, *prob*) 数值型函数, 函数值是一个来自具有指定实验次数  $n$  和概率参数 *prob* 的二项式分布的随机数。

二项分布要求试验成功的概率不能太小, 不能接近 0, 例如  $< 0.01$ 。如果事件的发生需要很大的样本量, 即  $n$  很大, 一次发生的概率  $p$  很小, 且  $np$  为常量时二项分布就趋近泊松分布了。

## (3) 泊松分布

如果某稀有事件的发生次数用随机变量  $X$  表示,  $X$  的取值范围是  $0, 1, 2, 3, \dots$ , 而且随机变量  $X = k$  的概率是

$$p(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad k = 0, 1, 2, 3, \dots, \infty \quad \lambda > 0$$

则称随机变量  $X$  服从参数为  $\lambda$  的泊松分布。式中,  $e$  是自然对数的底,  $e = 2.71828 \dots$

与泊松分布有关的函数:

- PDF.POISSON(*quant*, *mean*) 数值型函数, 函数值是具有指定均值和概率参数的泊松分布, 值等于 *quant* 的概率。
- CDF.POISSON(*quant*, *mean*) 数值型函数, 函数值是具有指定均值或概率参数的泊松分布, 随机变量值小于或等于 *quant* 的累积概率。
- RV.POISSON(*mean*) 数值型函数, 函数值是一个具有指定均值 *mean* 或比率 *rate* 参数的泊松分布的随机数。

3. 离散型随机变量的均值与标准差

设离散型随机变量  $X$  的概率分布是：

$X$	$x_1$	$x_2$	$\cdots$	$x_k$	$\cdots$
$P$	$p_1$	$p_2$	$\cdots$	$p_k$	$\cdots$

即  $P(X=x_k)=p_k, k=1, 2, \cdots$ ，则称和数

$$\sum_k x_k p_k$$

为随机变量  $X$  的期望，记作  $E(X)$ ，也称它为  $X$  的均值。同时，称和数

$$\sum_k [x_k - E(X)]^2 p_k$$

为随机变量  $X$  的方差，记作  $D(X)$ ，而方差的算术平方根，则称为标准差。

4. 连续型随机变量的概率分布

前面已经提到，连续型随机变量可以在某个定义的区间内，也可在某些区间中取任意实数。度量这些量的单位在理论上是可以无限再分的。

连续型随机变量取任何值的概率都是 0，只有在某个区间中的概率才可能不是 0。所以不能像离散型随机变量那样列出每一个值的相应概率。对连续型随机变量，我们用密度函数形式来描述。连续型概率密度函数  $f(x)$  满足下列两个条件

$f(x) \geq 0$

①

$\int_{-\infty}^{+\infty} f(x)dx = 1$

②

与离散型概率不同的是， $f(x)$  不是概率，而是概率密度函数。累积分布函数是连续分布的随机变量  $X$  小于某值  $x$  的概率  $P$ ，即  $P(X \leq x)$  以概率密度函数曲线在该区间的面积表示，即

$$F(x) = \int_{-\infty}^x f(x)dx$$

当  $x = a$  时，有概率

$F(a) = \int_{-\infty}^a f(x)dx$

当  $x = b$  时，有概率

$F(b) = \int_{-\infty}^b f(x)dx$

随机变量在某区间上的概率是上述公式②在某一个区间的积分。表示  $x$  值落在这个区间中的概率。

例如，连续随机变量  $x$  落在  $(a, b)$  区间中的概率(见图 4-1)是

$$P(a < x < b) = \int_a^b f(x)dx$$

那么，就有  $P(a < x < b) = F(b) - F(a)$

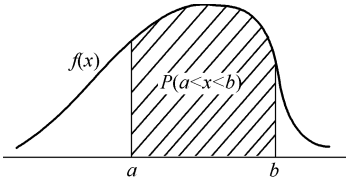


图 4-1 概率密度函数与概率

5. 连续型随机变量的均值与标准差

连续型随机变量的均值定义为：

$$\mu = \int_{-\infty}^{+\infty} xf(x)dx$$

连续型随机变量的标准差定义为：

$$\sigma = \sqrt{\int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx}$$

常用连续型随机变量的概率分布如下。

(1) 指数分布的概率密度函数为：

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

与指数分布有关的 SPSS 函数为：

- PDF.EXP(*quant*, *shape*) 数值型函数，函数值为形状参数为 *shape* 的指数分布在 *quant* 处的概率密度。
- CDF.EXP(*quant*, *shape*) 数值型函数，函数值是具有给定的形状参数 *shape* 的指数分布，随机变量的值小于 *quant* 的累积概率。
- RVEXP(*shape*) 数值型函数，函数值是一个来自具有指定形状参数的指数分布的随机数。

(2) 正态分布的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < +\infty, \sigma > 0$$

式中， $\mu$  为随机变量  $X$  的均值， $\sigma$  为标准差，均为常数。随机变量服从均值为  $\mu$ 、标准差为  $\sigma$  的正态分布，记做  $X \sim N(\mu, \sigma)$ 。

当均值 0，标准差为 1 时的正态分布为标准正态分布，记做  $z \sim N(0, 1)$ 。

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

可以通过  $z$  变换实现随机变量的标准化：

$$z = \frac{x - \mu}{\sigma}$$

与正态分布函数有关的 SPSS 函数为：

- PDF.NORMAL(*quant*, *mean*, *stddev*) 数值型函数，函数值是具有指定均值和标准差的正态分布在 *quant* 处的概率密度。
- CDF.NORMAL(*quant*, *mean*, *stddev*) 数值型函数，返回一个均值为 *mean*，标准差为 *stddev* 的正态分布的随机变量值小于 *quant* 的累积概率。
- RV.NORMAL(*mean*, *stddev*) 数值型函数，函数值是一个具有指定均值 *mean* 和标准差 *stddev* 的正态分布随机数。

#### 4.1.2 随机变量的函数

SPSS 中随机变量的函数包括 7 类，概述于表 4-1 中。

随机变量和分布函数的关键字分前缀、后缀，前、后缀之间用圆点分隔。前缀指定分布的函数归类，后缀指定分布。

随机变量和分布函数的自变量可以是常量，也可以是变量。

如果要求函数自变量，对累积分布函数、概率密度函数和反分布函数的概率  $p$ ，必须出现在第一个，用  $x$  表示(*quant* 必须落在分布的合法值范围内)。

对随机变量和分布函数，必须指定分布参数作为对分布的说明，所有自变量都是实数。

表 4-1 7 类随机变量函数概述

类	解 释	数 目
CDF	累积分布函数 $CDF.d\_spec(x,a,...)$ 其值是累积概率 $p$ , 具有指定的 ( $d\_spec$ ) 分布的连续型随机变量落在 $x$ 以下的累积概率; 对离散型随机变量来说是在 $x$ 处或 $x$ 以下的概率	26
IDF	反分布函数对离散分布不能用。 反分布函数 $IDF.d\_spec(p,a,...)$ 的函数值是 $CDF.d\_spec(x,a,...)=p$ 的具有 ( $d\_spec$ ) 指定分布的 $x$ 值	18
PDF	概率密度函数 $PDF.d\_spec(x,a,...)$ 其值对连续型随机变量来说是指定分布在 $x$ 处的概率密度, 对离散型随机变量来说是具有指定分布的随机变量值等于 $x$ 的概率	23
RV	随机数发生函数 $RV.d\_spec(a,...)$ 产生独立的具有指定分布 ( $d\_spec$ ) 的观测	22
NCDF	非中心累积分布函数 $NCDF.d\_spec(x,a,b,...)$ 的值是一个具有指定的非中心分布的变量落在 $x$ 以下的概率 $p$ , 只对贝塔 ( $\beta$ ) 分布、卡方分布、 $F$ 分布和学生 $T$ 分布可用	4
NPDF	非中心概率密度函数 $NCDF.d\_spec(x,a,b,...)$ 的值是具有指定分布 ( $d\_spec$ ) 的随机变量在 $x$ 处的概率密度, 只对贝塔 ( $\beta$ ) 分布、卡方分布、 $F$ 分布和学生 $T$ 分布可用	4
SIG	显著性函数 $SIG.d\_spec(x,a,...)$ 的值是具有指定分布 ( $d\_spec$ ) 的变量大于 $x$ 的概率 $p$ , 它等于 1 减去累积分布函数值	2

注意: SPSS 20.0 版本把累积分布函数 CDF 与非中心累积分布函数 NCDF 归为一类; 把概率密度函数与非中心概率密度函数归为一类。提醒读者在计算变量或其他应用而查找这些函数时注意。

1. 随机数函数(Random Numbers, 22 个)

下面的函数根据指定的分布给出一个随机变量值, 自变量是分布参数。

如果在数据文件中建立新变量时使用这些函数, 则变量值的个数等于数据文件中合法的观测数。

注意: 函数名中的圆点是半角的圆点。

也可以事先在【随机数字生成器】对话框中通过设置一个种子, 再由循环结构程序产生一系列符合一定分布的伪随机数。按【转换→随机数字生成器】顺序单击菜单项, 打开如图 4-2 所示的对话框。在对话框中有以下两栏选项。

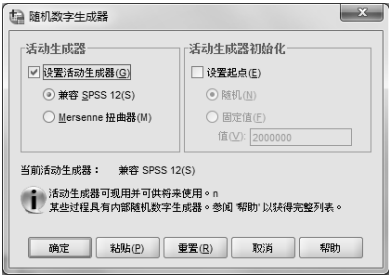


图 4-2 【随机数字生成器】对话框

①【活动生成器】栏。有两个随机数字生成器供选择。选择【设置活动生成器】复选项。用户设置工作生成器可选择:

- 【兼容 SPSS 12】的生成器。如果需要再生成利用 12 版本或 12 版本以前的生成器, 基于一个种子值的随机化结果, 就选择这个生成器。
- 【Merscenne 扭曲器】。一个新的更可靠的随机数生成器。

②【活动生成器初始化】栏。设置现行生成器初始值。选择【设置起点】复选项可以由读者设置随机数字生成器的初始种子值。有两个选项:

- 【随机】。即由系统给出随机数作为产生随机数的种子, 系统默认此选项。
- 【固定值】。由用户设定。在【值】后面设置一个数值。

随机变量函数如下:

(1)  $RV.BERNOULLI(prob)$  数值型函数。函数值为一个来自伯努利分布且具有指定概率参数  $prob$  的随机数。

(2)  $RV.BETA(shape1, shape2)$  数值型函数。函数值是一个来自具有指定形状参数  $shape1$ ,  $shape2$  的 Beta 分布的随机数。

(3)  $\text{RV.BINOM}(n, prob)$  数值型函数。函数值是一个来自具有指定实验次数  $n$  和概率参数  $prob$  的二项式分布的随机数。

(4)  $\text{RV.CAUCHY}(loc, scale)$  数值型函数。函数值是一个来自具有指定位置  $loc$  和尺度  $scale$  参数的柯西分布的随机数。

(5)  $\text{RV.CHISQ}(df)$  数值型函数。函数值是一个来自具有指定自由度  $df$  的卡方分布的随机数。

(6)  $\text{RV.EXP}(shape)$  数值型函数。函数值是一个来自具有指定形状参数  $shape$  的指数分布的随机数。

(7)  $\text{RV.F}(df1, df2)$  数值型函数。函数值是一个来自具有指定自由度  $df1$ 、 $df2$  的 F 分布的随机数。

(8)  $\text{RV.GAMMA}(shape, scale)$  数值型函数。函数值是一个来自具有指定形状  $shape$  和尺度  $scale$  参数的伽马分布的随机数。

(9)  $\text{RV.GEOM}(prob)$  数值型函数。函数值是一个来自具有指定概率参数  $prob$  的几何分布的随机数。

(10)  $\text{RV.HALFNRM}(mean, stddev)$  数值型函数。函数值是一个具有指定均值  $mean$ 、标准差  $stddev$  的半正态分布的随机数。

(11)  $\text{RV.HYPER}(total, sample, hits)$  数值型函数。函数值是一个来自具有指定参数的超几何分布的随机数。

(12)  $\text{RV.IGAUSS}(loc, scale)$  数值型函数，函数值是一个来自具有指定位置参数  $loc$  和尺度参数  $scale$  的逆高斯分布的随机数。

(13)  $\text{RV.LAPLACE}(mean, scale)$  数值型函数。函数值是一个来自具有指定均数  $mean$  和尺度  $scale$  参数的拉普拉斯分布的随机数。

(14)  $\text{RV.LNORMAL}(a, b)$  数值型函数。函数值是一个来自具有指定参数  $a, b$  的对数正态分布随机数。

(15)  $\text{RV.LOGISTIC}(mean, scale)$  数值型函数。函数值是一个来自具有指定均数  $mean$  和尺度参数  $scale$  的 Logistic 分布的随机数。

(16)  $\text{RV.NEGBIN}(threshold, prob)$  数值型函数。函数值是一个具有指定阈值  $threshold$  和概率  $prob$  参数的负二项分布随机数。

(17)  $\text{RV.NORMAL}(mean, stddev)$  数值型函数。函数值是一个具有指定均值  $mean$  和标准差  $stddev$  的正态分布的随机数。

(18)  $\text{RV.PARETO}(threshold, shape)$  数值型函数。函数值是一个具有指定阈值  $threshold$  和形状参数  $shape$  的帕累托分布的随机数。

(19)  $\text{RV.POISSON}(mean)$  数值型函数。函数值是一个具有指定均值  $mean$  或比率  $rate$  参数的泊松分布的随机数。

(20)  $\text{RV.T}(df)$  数值型函数。函数值是一个来自具有指定自由度  $df$  的学生  $T$  分布的随机数。

(21)  $\text{RV.UNIFORM}(min, max)$  数值型函数。函数值是一个属于具有指定最大值  $max$  和最小值  $min$  的均匀一致分布的随机数，另请参考  $\text{UNIFORM}$  函数。

(22)  $\text{RV.WEIBULL}(a, b)$  数值型函数。函数值是一个属于具有指定参数  $a, b$  的威布尔分布的随机数。

## 2. 概率密度函数 PDF 与非中心 PDF(即 NPDF)(27 个)

下列函数给出具有指定分布, 在第一个自变量 *quant* 值处的密度函数的值, 后面的自变量是分布参数。

注意: 每个函数名中的句点是英文半角的。

(1) PDF.BERNOULLI(*quant*, *prob*) 数值型函数。函数值等于分布参数为 *prob* 的伯努利分布在 *quant* 处的概率值。

(2) PDF.BETA(*quant*, *shape1*, *shape2*) 数值型函数。函数值等于形状参数为 *shape1*、*shape2* 的 Beta 分布, 在 *quant* 处的概率密度值。

(3) PDF.BINOM(*quant*, *n*, *prob*) 数值型函数。每次实验成功的概率是 *prob* 时, 函数值为 *n* 次实验中的成功次数等于 *quant* 的概率。当 *n*=1 时, 该函数与 PDF.BERNOULLI 相同。

(4) PDF.BVNOR(*quant1*, *quant2*, *corr*) 数值型函数。函数值为具有给定相关系数 *corr* 的标准二元正态分布, 在 *quant1*, *quant2* 处的概率密度值。

(5) PDF.CAUCHY(*quant*, *loc*, *scale*) 数值型函数。函数值为具有给定位置参数 *loc* 和尺度参数 *scale* 的 Cauchy 分布在 *quant* 处的概率密度。

(6) PDF.CHISQ(*quant*, *df*) 数值型函数。函数值为自由度为 *df* 的卡方分布在 *quant* 处的概率密度。

(7) PDF.EXP(*quant*, *shape*) 数值型函数。函数值为形状参数为 *shape* 的指数分布在 *quant* 处的概率密度。

(8) PDF.F(*quant*, *df1*, *df2*) 数值型函数。函数值为自由度为 *df1*、*df2* 的 F 分布在 *quant* 处的概率密度。

(9) PDF.GAMMA(*quant*, *shape*, *scale*) 数值型函数。返回形状参数为 *shape*, 尺度参数为 *scale* 的 Gamma 分布在 *quant* 处的概率密度。

(10) PDF.GEOM(*quant*, *prob*) 数值型函数。返回概率值是当成功的概率是给定的 *prob* 时获得成功的实验数等于 *quant* 的概率。

(11) PDF.HALFNRM(*quant*, *mean*, *stddev*) 数值型函数。返回均值为 *mean*, 标准差为 *stddev* 的半正态分布在 *quant* 处的概率密度。

(12) PDF.HYPER(*quant*, *total*, *sample*, *hits*) 数值型函数。返回的数值是, 当从大小为 *total* 的, 具有指定特征的总体的 *hits* 个对象中随机选取样本 *sample* 时, 采样数中具有指定特征的对象数等于 *quant* 的概率。

(13) PDF.IGAUSS(*quant*, *loc*, *scale*) 数值型函数。返回具有给定的位置参数 *loc* 和尺度参数 *scale* 的逆高斯分布, 在 *quant* 处的概率密度。

(14) PDF.LAPLACE(*quant*, *mean*, *scale*) 数值型函数。返回具有指定均值 *mean* 和尺度参数 *scale* 的拉普拉斯分布, 在 *quant* 处的概率密度值。

(15) PDF.LNORMAL(*quant*, *a*, *b*) 数值型函数。函数值是具有指定参数 *a*、*b* 的对数正态分布, 在 *quant* 处的概率密度值。

(16) PDF.LOGISTIC(*quant*, *mean*, *scale*) 数值型函数。返回具有指定均值 *mean* 和尺度参数 *scale* 的 Logistic 分布, 在 *quant* 处的概率密度值。

(17) PDF.NEGBIN(*quant*, *thresh*, *prob*) 数值型函数。函数值是当阈值参数是给定的 *thresh*, 成功的概率是 *prob* 时, 获得一次成功的实验数等于 *quant* 的概率。



(18) PDF.NORMAL(*quant*, *mean*, *stddev*) 数值型函数, 函数值是具有指定的均值和标准差的正态分布, 在 *quant* 处的概率密度。

(19) PDF.PARETO(*quant*, *threshold*, *shape*) 数值型函数。函数值是具有指定阈值 *threshold* 和形状参数 *shape* 的帕累托分布, 在 *quant* 处的概率密度。

(20) PDF.POISSON(*quant*, *mean*) 数值型函数。函数值是具有指定均值和概率参数的泊松分布, 值等于 *quant* 的概率。

(21) PDF.T(*quant*, *df*) 数值型函数。函数值是具有指定自由度 *df* 的学生 *T* 分布, 在 *quant* 处的概率密度。

(22) PDF.UNIFORM(*quant*, *min*, *max*) 数值型函数。函数值是具有指定的最小值 *min* 参数和最大值参数 *max* 的一致分布, 在 *quant* 处的概率密度。

(23) PDF.WEIBULL(*quant*, *a*, *b*) 数值型函数, 函数值是具有指定参数 *a*、*b* 的威布尔分布在 *quant* 处的概率密度。

(24) NPDF.BETA(*quant*, *shape1*, *shape2*, *nc*) 数值型函数。函数值是具有给定形状参数 *shape1*、*shape2* 和非中心参数 *nc* 的非中心 beta 分布在 *quant* 处的概率密度。

(25) NPDF.CHSQ(*quant*, *df*, *nc*) 数值型函数。函数值是具有这样的 *df* 和非中心参数 *nc* 的非中心卡方分布在 *quant* 处的概率密度。

(26) NPDF.F(*quant*, *df1*, *df2*, *nc*) 数值型函数。函数值是具有自由度 *df1*、*df2* 和非中心参数 *nc* 的非中心 *F* 分布在 *quant* 处的概率密度。

(27) NPDF.T(*quant*, *df*, *nc*) 数值型函数。函数值是具有自由度 *df1*、*df2*, 且非中心参数为 *nc* 的非中心 *T* 分布在 *quant* 处的概率密度。

以上 (24) ~ (27) 为非中心分布的概率密度函数。

### 3. 累积分布函数 CDF 与非中心累积分布函数 NCDF (30 个)

下面的函数给出具有指定分布参数的随机变量值小于第一个自变量 *quant* 的累积概率, 分布类型由函数名决定, 后面的自变量是分布参数。

**注意:** 函数名中的圆点必须是英文半角圆点。

(1) CDF.BERNOULLI(*quant*, *prob*) 数值型函数。给出符合概率为 *prob* 的二项分布的随机变量, 其值小于等于 *quant* 的累积概率值。

(2) CDF.BETA(*quant*, *shape1*, *shape2*) 数值型函数。函数值为具有给定形状参数 *shape1*、*shape2* 的 Bate 分布的随机变量, 值小于 *quant* 的累积概率。

(3) CDF.BINOM(*quant*, *n*, *prob*) 数值型函数。当每次实验成功的概率是 *prob* 时, 函数值是一个 *n* 次实验中成功次数小于或等于 *quant* 的二项分布累积概率值; 当 *n*=1 时, 该函数与 CDF.BERNOULLI 相同。

(4) CDF.BVNOR(*quant1*, *quant2*, *corr*) 数值型函数。给出的函数值为来自二元标准正态分布的两个随机变量, 相关系数为 *corr*, 这两个随机变量值分别小于 *quant1*、*quant2* 的累计概率。

(5) CDF.CAUCHY(*quant*, *loc*, *scale*) 数值型函数。函数值是具有给定的位置参数 *loc* 和尺度参数 *scale* 的柯西分布的随机变量, 其值小于 *quant* 的累积概率值。

(6) CDF.CHISQ(*quant*, *df*) 数值型函数。函数值是具有给定自由度 *df* 的卡方分布的随机变量, 其值小于 *quant* 累积概率。

(7) CDF.EXP(*quant*, *shape*) 数值型函数。函数值是具有给定形状参数 *shape* 的指数分布的随机变量, 其值小于 *quant* 累积概率。

(8) CDF.F(*quant*, *df1*, *df2*) 数值型函数。函数值是具有给定自由度 *df1*、*df2* 的 *F* 分布的随机变量, 其值小于 *quant* 的累积概率值。

(9) CDF.GAMMA(*quant*, *shape*, *scale*) 数值型函数。函数值是具有给定形状参数 *shape* 和尺度参数 *scale* 的伽马分布的随机变量, 其值小于 *quant* 的累积概率。

(10) CDF.GEOM(*quant*, *prob*) 数值型函数。函数值是成功概率为 *prob* 的几何分布获得一次成功的实验次数。

(11) CDF.HALFNRM(*quant*, *mean*, *stddev*) 数值型函数。函数值是具有指定均值 *mean*, 标准差 *stddev* 的半正态分布的随机变量, 其值小于 *quant* 的累积概率值。

(12) CDF.HYPER(*quant*, *total*, *sample*, *hits*) 数值型函数。样品 *sample* 个事件是从大小为 *total* 的有 *hits* 个具有指定特性的总体中随机选择出来的情况下, 返回随机变量小于或等于 *quant* 的累积概率, 即具有指定特性的事件数。

(13) CDF.IGAUSS(*quant*, *loc*, *scale*) 数值型函数。函数值为具有给定的位置参数 *loc* 和尺度参数 *scale* 的逆高斯分布的随机变量, 其值小于 *quant* 的累积概率。

(14) CDF.LAPLACE(*quant*, *mean*, *scale*) 数值型函数。返回来自均值为 *mean*, 尺度参数为 *scale* 的拉普拉斯分布的随机变量, 值小于 *quant* 的累积概率。

(15) CDF.LNORMAL(*quant*, *a*, *b*) 数值型函数。返回具有指定参数 *a*、*b* 的对数正态分布的随机变量值小于 *quant* 的累积概率。

(16) CDF.LOGISTIC(*quant*, *mean*, *scale*) 数值型函数。返回来自具有给定的均值 *mean* 和尺度参数 *scale* 的 Logistic 分布的随机变量值小于 *quant* 的累积概率。

(17) CDF.NEGBIN(*quant*, *thresh*, *prob*) 数值型函数。即当阈值参数为 *thresh*, 成功的概率为 *prob* 时, 在 *quant* 次实验中获得一次成功的实验次数。

(18) CDF.NORMAL(*quant*, *mean*, *stddev*) 数值型函数。返回一个均值为 *mean*, 标准差为 *stddev* 的正态分布的随机变量值小于 *quant* 的累积概率。

(19) CDF.PARETO(*quant*, *threshold*, *shape*) 数值型函数。返回阈值为 *threshold*, 形状参数为 *shape* 的帕累托分布的随机变量值小于 *quant* 的累积概率。

(20) CDF.POISSON(*quant*, *mean*) 数值型函数。函数值是具有指定的均值或概率参数的泊松分布的随机变量值小于 *quant* 的累积概率。

(21) CDF.SMOD(*quant*, *a*, *b*) 数值型函数。返回具有指定参数 *a*、*b*, 属于学生化的最大模的随机变量值小于 *quant* 的累积概率。

(22) CDF.SRANGE(*quant*, *a*, *b*) 数值型函数。函数值是具有指定参数 *a*、*b* 的学生化值域分布的随机变量值小于 *quant* 的累积概率。

(23) CDF.T(*quant*, *df*) 数值型函数。函数值是具有指定自由度 *df* 的学生 *T* 分布的随机变量, 值小于 *quant* 的累积概率。

(24) CDF.UNIFORM(*quant*, *min*, *max*) 数值型函数。函数值是具有指定的最小值 *min* 和最大值 *max* 参数的一致分布的随机变量, 值小于 *quant* 的累积概率。

(25) CDF.WEIBULL(*quant*, *a*, *b*) 数值型函数。函数值是具有指定的参数 *a*、*b* 的威布尔分布的随机变量, 值小于 *quant* 的累积概率。

(26) CDF.NORM(*zvalue*) 数值型函数。返回一个均值为 0, 标准差为 1 的标准正态分布的随机变量, 值小于 *zvalue* 的概率。

(27) NCDF.BETA(*quant*, *shape1*, *shape2*, *nc*) 数值型函数。返回一个具有指定的形状参数

$shape1$ 、 $shape2$  和非中心参数  $nc$  的 Beta 分布的随机变量, 值小于  $quant$  的累积概率。

(28) NCDF.CHISQ( $quant, df, nc$ ) 数值型函数。返回一个具有指定的自由度  $df$ 、非中心性参数  $nc$  的无偏卡方分布的随机变量, 值小于  $quant$  的累积概率。

(29) NCDF.F( $quant, df1, df2, nc$ ) 数值型函数。返回一个具有指定的自由度  $df1$ 、 $df2$  和非中心性参数  $nc$  的非中心  $F$  分布的随机变量, 值小于  $quant$  的累积概率。

(30) NCDF.T( $quant, df, nc$ ) 数值型函数。返回一个具有指定的自由度  $df$ 、非中心性参数  $nc$  的非中心  $T$  分布的随机变量, 值小于  $quant$  的累积概率。

#### 4. 反分布函数(Inverse DF, 18 个)

下面的函数给出一个在指定的分布中的值, 这个分布的累积概率为第一个自变量  $prob$  的值, 其后的自变量是指定分布的参数。注意, 每个函数名由两部分组成, 圆点前是函数类名, 圆点后是分布名称, 括号内是自变量。当已知某分布累积概率值, 求随机变量值时使用此类函数。

(1) IDF.BETA( $prob, shape1, shape2$ ) 数值型函数。函数值为形状参数为  $shape1$ 、 $shape2$  的 Beta 分布的随机变量, 在累积概率为  $prob$  处的值。

(2) IDF.CAUCHY( $prob, loc, scale$ ) 数值型函数。函数值为位置参数  $loc$  和尺度参数  $scale$  的柯西分布的随机变量, 在累积概率为  $prob$  处的值。

(3) IDF.CHISQ( $prob, df$ ) 数值型函数。函数值为一个卡方值, 该卡方分布的自由度为  $df$ , 概率值为  $prob$ 。例如, 在 0.05 水平上(累积概率为 95%), 自由度为 3 的卡方值为 IDF.CHISQ(0.95,3)。

(4) IDF.EXP( $p, scale$ ) 数值型函数。函数值为按  $scale$  速度指数衰减的随机变量, 累积概率在  $p$  处的值。

(5) IDF.F( $prob, df1, df2$ ) 数值型函数。函数值为自由度为  $df1$ 、 $df2$  的  $F$  分布的随机变量累积概率为  $prob$  的值。例如, 显著性概率在 0.05 水平上, 自由度分别为 3、100 的  $F$  值为 IDF.F(0.95,3,100)。

(6) IDF.GAMMA( $prob, shape, scale$ ) 数值型函数。函数值为形状参数为  $shape$  和尺度参数为  $scale$  的伽马分布的随机变量, 累积概率为  $prob$  的值。

(7) IDF.HALFNRM( $prob, mean, stddev$ ) 数值型函数。函数值为一个具有指定的均值  $mean$  标准差  $stddev$  的半正态分布的随机变量, 累积概率为  $prob$  的值。

(8) IDF.IGAUSS( $prob, loc, scale$ ) 数值型函数。函数值为具有给出的位置参数  $loc$  和尺度参数  $scale$  的逆高斯分布随机变量, 累积概率是  $prob$  的值。

(9) IDF.LAPLACE( $prob, mean, scale$ ) 数值型函数。函数值等于均值为  $mean$  和尺度参数为  $scale$  的拉普拉斯分布的随机变量, 累积概率为  $prob$  的值。

(10) IDF.LNORMAL( $prob, a, b$ ) 数值型函数。函数值为有指定参数  $a, b$  的对数正态分布的随机变量, 累积概率为  $prob$  的值。

(11) IDF.LOGISTIC( $prob, mean, scale$ ) 数值型函数。函数值等于均值为  $mean$  和尺度参数为  $scale$  的 Logistic 分布的随机变量, 累积概率为  $prob$  的值。

(12) IDF.NORMAL( $prob, mean, stddev$ ) 数值型函数。函数值为具有指定均值和标准差的正态分布随机变量, 累积概率为  $prob$  的值。

(13) IDF.PARETO( $prob, threshold, shape$ ) 数值型函数。函数值为阈值  $threshold$ , 尺度参数  $scale$  的帕累托分布的随机变量, 在累积概率  $prob$  处的值。

(14)  $\text{IDF.SMOD}(prob, a, b)$  数值型函数。函数值是具有指定参数  $a, b$  的学生最大模数随机变量, 累积概率是  $prob$  的值。

(15)  $\text{IDF.SRANGE}(prob, a, b)$  数值型函数。函数值为具有指定参数  $a, b$  的学生化范围统计量, 累积概率是  $prob$  的值。

(16)  $\text{IDF.T}(prob, df)$  数值型函数。函数值为具有指定自由度  $df$  的学生  $T$  分布的随机变量, 其累积概率为  $prob$  的值。

(17)  $\text{IDF.UNIFORM}(prob, min, max)$  数值型函数。函数值为具有的最大值  $max$ 、最小值  $min$  的均匀分布的随机变量, 其累积概率为  $prob$  的值。

(18)  $\text{IDF.WEIBULL}(prob, a, b)$  数值型函数。函数值是具有指定参数  $a, b$  的韦伯分布的随机变量累积概率为  $prob$  的值。

## 5. 显著性函数(Significance, 2 个)

下列函数给出具有指定分布的随机变量大于第一个自变量  $quant$  的概率, 后边的自变量是分布参数。

(1)  $\text{SIG.CHISQ}(quant, df)$  数值型函数。函数值是具有  $df$  自由度的卡方分布, 大于自变量  $quant$  值的累积概率。

(2)  $\text{SIG.F}(quant, df1, df2)$  数值型函数。函数值是自由度为  $df1, df2$  的  $F$  分布的, 大于自变量  $quant$  值的累积概率。

## 4.2 随机变量与分布函数的应用

### 4.2.1 符合分布要求的随机数的生成

**【例 1】** 如要生成均值为 0、标准差为 1 的正态分布的随机数 1000 个, 方法如下:

(1) 首先在数据编辑窗中输入序号变量  $no$ , 数值型变量, 输入编号 1~1000, 即设法制造 1000 个观测。

例如, 可以在 Excel 中生成 1~1000 的 1000 个数字, 然后选择并复制到剪贴板; 在 SPSS 中打开一个空数据文件, 建立编码变量  $no$ ; 将剪贴板中数据粘贴到该变量列中。

(2) 按【转换→计算变量】顺序单击菜单项, 打开【计算变量】对话框。

(3) 输入新变量名  $RNorm$ , 单击【类型与标签】按钮, 打开【计算变量: 类型和标签】对话框, 见图 4-3(a), 用中文填写标签。变量类型为数值型。单击【继续】按钮返回主对话框。

(4) 在【计算变量】对话框中的【函数组】栏中选择【随机数字】类。在【函数和特殊变量】栏中选择【 $Rv.Normal$ 】函数, 单击向上箭头按钮, 将该函数原型  $RV.NORMAL(?,?)$  显示在【数字表达式】栏内。输入函数参数: 均值 0、标准差 1 代替两个问号为  $RV.NORMAL(0,1)$ , 见图 4-3(b)。

(5) 单击【确定】按钮, 生成均值为 0、标准差为 1 的正态分布随机数, 显示在数据窗中。生成的部分数据见图 4-4(a)。

根据生成的正态随机数绘制直方图:

(1) 按【图形→旧对话框→直方图】顺序单击菜单项, 见图 4-4, 打开【直方图】对话框, 见图 4-5。

(2) 选择随机变量  $RNorm$ , 单击向右箭头按钮, 将其移入【变量】栏内, 选择【显示正态曲线】复选项。要求在输出的直方图上同时显示标准正态曲线, 见图 4-5(a)。



图 4-3 生成正态分布的随机数操作过程示意图

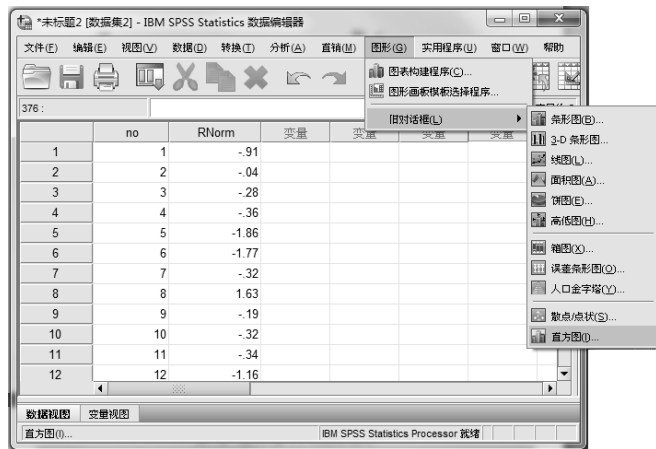


图 4-4 生成的正态随机数据、绘制直方图菜单



图 4-5 【直方图】主对话框和【标题】对话框

(3) 单击【标题】按钮打开如图 4-5(b)所示的【标题】对话框。在第一行中输入“正态分布随机数发生函数验证”。单击【继续】按钮，返回主对话框。

单击【确定】按钮，生成如图 4-6 所示的直方图，实线为标准正态曲线。

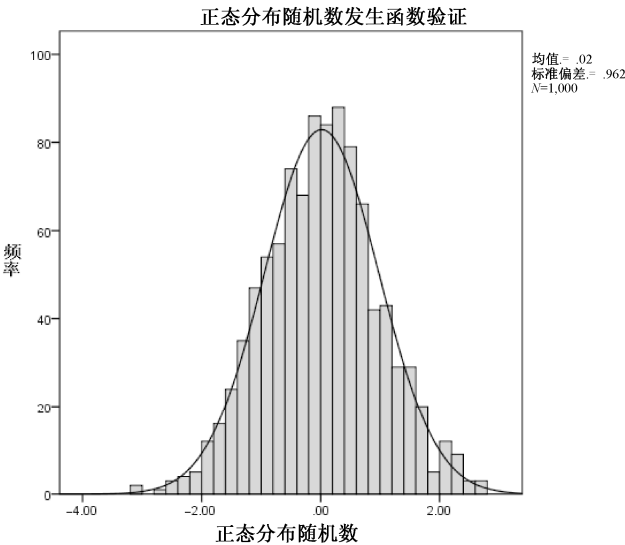


图 4-6 标准正态随机数绘制出的直方图

从图中可以看出，随机数函数产生的数据的确是正态分布的。  
本例给出了一个从感性认识一种分布的方法。在学习统计学以及教学中是很有用的。

4.2.2 概率密度函数与累积概率密度函数的应用

【例 2】 离散随机变量及其分布的应用。

某体育专科学校改革课题的调查表明，该类学校 75%的教师认为学生严重缺乏应该在中学阶段就掌握的基本技能。假定该校同意这一看法的总体比例是  $\pi=0.75$ ，在某校抽取 20 名教师组成样本，问：20 人中有 11 人同意该意见的概率有多大？小于等于 11 人同意该意见的概率有多大？多于 11 人同意该意见的概率有多大？

解：同意与否是两个互斥事件，本例中的实验结果数据属于二项分布。设同意该意见的人数为  $x$ 。原题意为求  $P(x=11)$ 、 $P(x \leq 11)$  的概率和  $P(x > 11)$  的概率。

使用 SPSS 的 PDF.BINOM 函数解决这个问题， $P=0.75$ ， $n$  次实验， $n=20$ ， $quant=11$ 。

- ① 建立一个变量  $x$ ，标签为“同意的人数”，输入数据 1~15。
- ② 建立另一个变量  $p$ ，标签为“同意的概率”输入数据为 15 个相同的数据 0.75，见图 4-7 (a)。
- ③ 顺序单击【转换→计算变量】，打开【计算变量】对话框，见图 4-7 (b)。
  - 在【目标变量】栏内输入新变量名“prob”。
  - 单击【类型与标签】按钮打开对话框，设置【类型】为数值型；设置【标签】为“20 人中  $x$  人同意的概率”。单击【继续】按钮，返回主对话框。
  - 在【函数组】栏中选择 PDF 与非中心 PDF 类，在【函数和特殊变量】栏中选择【PDF.Binom】函数，将函数送入【数字表达式】栏内。
  - 在函数名后的括号中按顺序输入：“ $x,20,p$ ”。函数显示为：PDF.BINOM( $x,20,p$ )
  - 单击【确定】按钮，在数据窗中生成变量  $prob$  的 15 个值，见图 4-8。



图 4-7 用 SPSS 函数求概率的方法

④ 20 人中有 11 人同意的概率可直接在数据窗中查  $x = 11$  时的  $prob$  值, 其为 0.027061。(在变量观察窗中将变量的小数位数增加为 5 或 6, 见数据文件 data04-01。)

使用 CDF.Binom 函数, 重复上述操作, 生成新变量  $Cprob$ , 【数字表达式】为

$$CDF.Binom(x, 20, p)$$

小于等于 11 人同意该意见的概率, 查  $x = 11$  的  $Cprob$  值, 即  $CDF.Binom(11, 20, 0.75)$ , 该值为 0.040925。

多于 11 人同意该意见的概率为  $1 - CDF.Binom(11, 20, 0.75) = 0.959075$ , 见图 4-9。

	x	p	prob
1	1	.75	.000000
2	2	.75	.000000
3	3	.75	.000000
4	4	.75	.000000
5	5	.75	.000003
6	6	.75	.000026
7	7	.75	.000154
8	8	.75	.000752
9	9	.75	.003007
10	10	.75	.009922
11	11	.75	.027061
12	12	.75	.060887
13	13	.75	.112406
14	14	.75	.168609
15	15	.75	.202331

图 4-8 可查询概率的数据窗

	x	p	prob	Cprob
1	1	.75	.000000	.000000
2	2	.75	.000000	.000000
3	3	.75	.000000	.000000
4	4	.75	.000000	.000000
5	5	.75	.000003	.000004
6	6	.75	.000026	.000030
7	7	.75	.000154	.000184
8	8	.75	.000752	.000935
9	9	.75	.003007	.003942
10	10	.75	.009922	.013864
11	11	.75	.027061	.040925
12	12	.75	.060887	.101812
13	13	.75	.112406	.214218
14	14	.75	.168609	.382827
15	15	.75	.202331	.585158
16	16	.75	.189685	.774844
17	17	.75	.133896	.908740
18	18	.75	.066948	.975687
19	19	.75	.021141	.996829
20	20	.75	.003171	1.000000

图 4-9 可查询概率和累积概率的数据窗

【例 3】连续随机变量及其分布的应用。

在市场调查中，顾客对折扣优惠有不同看法和态度。一项调查对使用和不使用折扣优惠券的某种品牌的饮料价格进行比较。得出平均差价为 5.5 角，标准差为 3.5 角。假定差价  $x$ (角)服从正态分布。求差价大于等于 10 角的概率、大于等于 5 角的概率，以及小于 0 角的概率。

注意：钱不是连续型的随机变量，这里只是介绍一种解决问题的思路与方法。


首先分析，大于等于 10 角的概率就是 1 减去差价小于 10 角的累积概率；同样，大于等于 5 角的概率就是 1 减去差价小于 5 角的累积概率。因此应该选用累积概率函数解决此问题。

- (1) 建立变量  $x$ ，数值型，输入数据 0~10。
- (2) 单击【转换→计算变量】，打开【计算变量】对话框，见图 4-10。

① 在【目标变量】栏输入变量名“ $C_p$ ”，单击【类型与标签】按钮，在打开的对话框中定义【变量类型】为数值型，【变量标签】为“累积概率”，见图 4-11。



图 4-10 求累积概率的过程——计算变量对话框

② 在【函数组】栏选择【CDF 与非中心 CDF】类，在【函数和特殊变量】栏内选择【CDF.Normal】函数，单击 ，将其移到【数字表达式】栏中。

在数据编辑窗口的变量观察窗口增加变量  $C_p$  的小数位数至 6 位。

③ 在函数自变量位置的 3 个问号处，分别输入“ $x$ ”、“5.5”、“3.5”，为  $Cdf.Normal(x,5.5,3.5)$ ，见图 4-7(b)，单击【确定】按钮。

④ 数据观察窗口计算结果见图 4-11，数据文件见 data04-02。

$x = 10$  时， $C_p = 0.90073$ ； $x = 5$  时， $C_p = 0.44320$ ； $x = 0$  时， $C_p = 0.05804$ ；因此差价大于等于 10 角(1 元)的概率为： $1 - 0.90073 = 0.09927$ ；差价大于等于 5 角(1 元)的概率为： $1 - 0.44320 = 0.5568$ ；差价小于 0 的概率为 0.05804。





图 4-11 计算结果

### 习 题 4

1. 某汽车公司的汽车销售量在过去 300 天的营业时间里，有 55 天销售量为 0；有 118 天销售量为 1；有 70 天为 2 辆；40 天为 3 辆；10 天为 4 辆；7 天为 5 辆。以过去 300 天的销售为历史数据，问：一天中售出 0、1、2、3、4、5 辆汽车的概率分别是多少？以此验证离散型随机变量的概念与性质。
2. 用 RV.BERNOULLI 函数生成符合伯努利分布、概率为 0.4 的 1000 个随机数；用 RV.LNORMAL 函数，生成参数  $a = 0.2$ 、 $b = 0.5$  的符合对数正态分布的随机数 1000 个，并作直方图。
3. 某仪器上的部件长度要求非常严格，要求在 0.304~0.322 cm 之间，某生产厂家生产的部件近似服从均值为 0.3015、标准差为 0.0016 的正态分布。求该生产厂家的不合格率。经改进，该厂产品近似服从均值为 0.3146、标准差为 0.0030，不合格率为多少？

# 第 5 章 日期和时间函数及其运算

## 5.1 日期时间函数

### 5.1.1 SPSS 日期时间概述

SPSS 的日期时间函数都借用固定数值进行转换，这个固定数值是 1582 年 10 月 14 日 0 时 0 分 0 秒。函数自变量无论是 *timevalue* 还是 *datevalue* 都是以这个日期时间为基准的。如果以一个数值型变量作日期时间函数的自变量，日期时间函数将自变量的值看作自 1582 年 10 月 14 日 0 时 0 分 0 秒算起的秒数。在 SPSS 中输入 1582/10/14 以前的日期，系统自动将其转换为缺失值。因此，如果把两个日期型变量直接进行运算时，要注意计算结果的类型和使用之后的函数进行转换后的数值所代表的含义。

### 5.1.2 日期时间常量与变量

#### 1. 日期常量

日期常量的表示方法很多。SPSS 为适应世界上不同国家、地区表示日期和时间的习惯，有多达 25 种表示方法，见表 5-1。示例中显示了可以直接使用的日期和时间的输入方法。中国人习惯的年、月、日顺序的表示方法，分为 2 位年和 4 位年两种，年、月、日之间用斜杠分隔，见表中灰色底纹的两行。

表 5-1 日期型常量格式及示例

格 式	说 明	示 例
dd-mmm-yyyy	日(2 位)-月份(英文)-年(4 位)	15-AUG-1945, 23-DEC-2008
dd-mmm-yy	日(2 位)-月份(英文)-年(2 位)	15-AUG-45, 23-DEC-95
mm/dd/yyyy	月份(2 位)/日(2 位)/年(4 位)	08/15/1945, 12/23/1995
mm/dd/yy	月份(2 位)/日(2 位)/年(2 位)	08/15/45, 12/23/95
dd.mm.yy yy	日(2 位).月份(英文).年(4 位)	08.15.1945, 12.23.95
dd.mm.yy	日(2 位).月份(英文).年(2 位)	08.15. 45
yyyy/mm/dd	年(4 位)/月(2 位)/日(2 位)	2008/07/07
yy/mm/dd	年(2 位)/月(2 位)/日(2 位)	08/08/15
yyddd	年(2 位)日数(自 1 月 1 日算起)	45227, 95
yyyyddd	年(4 位)日数(自 1 月 1 日算起)	1945227, 1995
q Q yyyy	季度 Q 年(4 位)	3Q1945, 4Q1995
q Q yy	季度 Q 年(2 位)	3Q45, 4Q95
mmm yyyy	月份(英文)年(4 位)	AUG1945, DEC1995
mmm yy	月份(英文)年(2 位)	AUG45, DEC95
ww WK yyyy	周数 WK 年(4 位)	33 WK 1945, 52 WK 1995

续表

格 式	说 明	示 例
ww WK yy	周数 WK 年(2 位)	33 WK 45, 52 WK 95
Monday, Tuesday...	直接输入英文的星期几	Friday
Mon, Tue, Wed...	直接输入星期几的英文缩写	FRI
January, February...	直接输入英文月份	August, December
Jan, Feb, Mar...	直接输入英文月份缩写	AUG, DEC
dd-mmm-yyyy hh:mm	日(2 位)-月(英文月份缩写)-年(4 位) 时(2 位):分(2 位)	11-AUG-1945 11:10
dd-mmm-yyyy hh:mm:ss	日(2 位)-月(英文月份缩写)-年(4 位) 时(2 位):分(2 位):秒(2 位)	11-AUG-1945 11:10:35
dd-mmm-yyyy hh:mm:ss.ss	日(2 位)-月(英文月份缩写)-年(4 位) 时(2 位):分(2 位):秒(2 位).百分秒	11-AUG-1945 11:10:35.30
hh:mm	时(2 位):分(2 位)	11:30, 08:50
hh:mm:ss	时(2 位):分(2 位):秒(2 位)	11:08:05, 08:15:25
hh:mm:ss.ss	时(2 位):分(2 位):秒(2 位).百分秒	11:08:05.80, 08:15:25.45
ddd hh:mm	日数 时(2 位):分(2 位)	128 08:50
ddd hh:mm:ss	日数 时(2 位):分(2 位):秒(2 位)	128 08:50:30
ddd hh:mm:ss.ss	日数 时(2 位):分(2 位):秒(2 位).百分秒	128 08:50:30.78

注：“m”在年与日(字母 y 与 d)之间表示“月”份，在时与秒(字母 h 与 s)之间表示“分”；“mmm”表示要求书写英文月份缩写；“ddd”表示要求用从 1 月 1 日算起的日数表示日期。

指定了日期型变量的格式，不一定在输入时就使用指定的格式输入。可以输入用“/”或“-”作分隔符的具体日期，按回车键后，系统自动将输入的日期转换为指定格式，显示在单元格中。

2. 日期时间变量

日期时间变量的输入/输出格式见表 5-2。

表 5-2 日期时间型变量输入/输出格式

格式类型	说 明	最小 w		最大 w	最大 d	一般格式	示 例
		输入	输出				
DATEw	国际通用	9	9	40		dd-mmm-yy	28-OCT-90
		10	11			dd-mmm-yyyy	28-OCT-1990
ADATEw	美国	8	8	40		mm/dd/yy	10/28/90
		10	10			mm/dd/yyyy	10/28/1990
EDATEw	欧洲	8	8	40		dd.mm.yy	28.10.90
		10	10			dd.mm.yyyy	28.10.1990
JDATEw	朱利安	5	5	40		yyddd	90301
		7	7			yyyyddd	1990301
SDATEw	可排序的日期*	8	8	40		yy/mm/dd	90/10/28
		10	10			yyyy/mm/dd	1990/10/28
QYRw	季度和年	4	6	40		q Q yy	4 Q 90
		6	8			q Q yyyy	4 Q 1990
MOYRw	月和年	6	6	40		mmm yy	OCT 90
		8	8			mmm yyyy	OCT 1990
WKYRw	星期和年	6	8	40		ww WK yy	43 WK 90
		8	10			ww WK yyyy	43 WK 1990
WKDAYw	一周的天	2	2	40		周内天的英文名	SU

续表

格式类型	说 明	最小 w		最大 w	最大 d	一般格式	示 例
		输入	输出				
MONTHw	月	3	3	40		月的英文名	JAN
TIMEw	时间	5	5	40		hh:mm	01:02
TIMEw.d		10	10	40	16	hh:mm:ss.s	01:02:34.75
DTIMEw	天数和时间	1	1	40		dd hh:mm	20 08:03
DTIMEw.d		13	13	40	16	dd hh:mm:ss.s	20 08:03:00
DATETIMEw	日期和时间	17	17	40		dd-mmm-yyyy hh:mm	20-JUN-1990 08:03
DATETIMEw.d		22	22	40	16	dd-mmm-yyyy hh:mm:ss.s	20-JUN-1990

合法日期或日期时间变量值无论是以什么格式输入的，转换成另一种日期或日期时间格式，都能正常显示。因为机内值是不变的，改变的只是输出(显示)格式。Date11(即 dd/mmm/yyyy)和 Date9(dd/mmm/yy)才是标准格式。有些函数只对标准格式有效。

5.1.3 日期时间函数

1. 当前日期时间函数(Current Date/Time)

(1) \$Date 字符型函数。其值为 9 位的 dd-mmm-yy 形式的当前日期。年数占 2 位。格式是 A9。字符型，要进行算术运算，必须转换成数值格式或日期格式。

(2) \$Date11 字符型函数。其值为 11 位的 dd-mmm-yyyy 的当前日期。年数占 4 位。格式 A11。字符型，要进行算术运算，必须转换为数值格式或日期格式。

(3) \$JDate，其值为数值型的当前日期，是用从 1582 年 10 月 14 日(罗马教皇格利高利定的第 1 天，即阳历第 1 天)算起的天数表示的当前日期。格式是 F6.0。

(4) \$Time，其值为当前日期和时间。\$TIME 给出的是从 1582 年 10 月 14 日 24:00:00 到转换命令执行之间的秒数。格式是 F20。可以把它显示为一个使用不同日期格式的数值的日期，也可以把它用在日期和时间函数中。

2. 日期的算术运算函数(Date Arithmetic)

(1) DATEDIFF(*datetime2*, *datetime1*, *unit*) 数值型函数。计算两个日期/时间值之间的差，并按指定的日期/时间单位 *unit* 返回一个整数(截去任何小数部分)。当 *datetime2* 和 *datetime1* 是日期或时间格式变量(或者是表示有效的日期时间的数值)，而 *unit* 是下列字符串之一：用引号括起的 years、quarters、months、weeks、days、hours、minutes、seconds。(年、季度、月、周、天、小时、分、秒)，表示差值转换后的时间单位。

(2) DATESUM(*datetime*, *value*, *unit*, *method*) 数值型函数。计算 *datetime* 指定的日期或时间格式的变量(或者表示合法日期/时间的数值)与日期时间值 *value* 之和，*unit* 是用引号括起的下列字符串值之一：years、quarters、months、weeks、days、hours、minutes、seconds(年、季度、月、周、天、小时、分、秒)，表示值 *value* 的单位。*method* 是可选的，可以是“rollover”或“closest”。

“rollover”：用滚动的方法把超出的天放到下一个月。  
“closest”：最近法，使用本月中最近的合法日期，这是默认的方法。

返回的值是表示成秒数的日期/时间值。要显示成日期/时间就要给变量赋予适当的格式。可以用转换函数将数值变量转换成日期/时间格式。

### 3. 时间生成函数(Date Creation)

这是一组数值型函数。函数值是将日期的某年、月、日、季度、周的数字的有效组合转变成自 1582 年 10 月 14 日 0 点 0 分 0 秒起至指定日期的秒数。自变量必须是整数。其中的 *year* 必须是 4 位大于 1582 的表示年的整数, *month* 是在 1~13 之间的月份。实际上有效值应该是 1~12, 如果输入数值为 13, 则按下一年的 1 月计算。 *quarter* 是在 1~4 之间的季度值, *weeknum* 是 1~52 之间的周数值, *daynum* 是在 1~366 之间的日数值。函数值是数值型, 要显示成日期, 只需在变量视图窗中将变量类型改变为日期型(Date 型变量)。

- (1) DATE.DMY(*day, month, year*) 数值型函数。返回与 *day, month* 和 *year* 相应的日期值。
- (2) DATE.MDY(*month, day, year*) 数值型函数。返回与 *month, day* 和 *year* 相应的日期值。
- (3) DATE.MOYR(*month, year*) 数值型函数。返回与 *month, year* 相应的日期值。
- (4) DATE.QYR(*quarter, year*) 数值型函数。返回与 *quarter, year* 相应的日期值。
- (5) DATE.WKYR(*weeknum, year*) 数值型函数。返回与 *weeknum, year* 相应的日期值。
- (6) DATE.YRDY(*year, daynum*) 数值型函数。返回与 *year, daynum* 相应的日期值。

### 4. 日期提取函数(Date Extraction)

在日期提取函数中, 需要注意:

① 这一组函数的自变量 *datevalue* (*timevalue*) 可以是:

- 数值或已经赋值的数值型变量或数值型表达式, 将自变量的值看作 1582 年 10 月 4 日 24:00:00 秒算起的天数(秒数)。
- 日期(时间)型变量、日期时间型表达式或日期、时间值, 机内值是从 1582 年 10 月 14 日 24:00:00 秒算起到达自变量指定日期(时间)的间隔中的天数(或秒数)。

② 自变量是 *timevalue* 的函数是数值型函数, 函数值为数值型常量。要把函数值显示成日期或时间, 应该赋予该函数值日期时间格式。

(1) XDATE.DATE(*datevalue*) 数值型函数。从数值返回表现为日期的日期部分。要把结果显示成日期, 需要把变量赋予日期格式。

实验表明, 因变量定义成数值型, 返回的是数值, 从 1582 年 10 月 14 日 0 点 0 分 0 秒到自变量指定的日期之间的秒数; 如果因变量定义成日期型, 函数值仍然是原日期不变。

(2) XDATE.JDAY(*datevalue*) 数值型函数。函数值为一年中的天数(1~366 之间的整数)。

(3) XDATE.MDAY(*datevalue*) 数值型函数。函数值是从 *datevalue* 提取出其代表日期中月份的第几天(在 1~31 之间的整数)。

(4) XDATE.MONTH(*datevalue*) 数值型函数。函数值是从 *datevalue* 提取出的月份(1~12 之间的整数)。

(5) XDATE.QUARTER(*datevalue*) 数值型函数。函数值是自变量所代表日期所在的一年中的季度(1~4 之间的整数)。

(6) XDATE.TDAY(*timevalue*) 数值型函数。从表现为时间间隔的自变量数值返回整数天数。

(7) XDATE.TIME(*datetime*) 数值型函数。从一个表现为时间或日期时间的值返回时间部分。要把结果显示成时间, 要赋予结果变量一个时间格式。

(8) XDATE.WEEK(*datevalue*) 数值型函数。函数值是自变量表达的日期在该年的周数(1~53 之间的整数)。

(9) XDATE.WKDAY(*datevalue*) 数值型函数。函数值为自变量 *datevalue* 表达的日期所在周中的天数(1 代表周日~7 代表周六之间的整数)。

(10) XDATE.YEAR(*datevalue*) 数值型函数。函数值是 4 位整数的年数。

(11) YRMODA(*year, month, day*) 数值型函数。根据自变量 *year*、*month*、*day* 返回从 1582 年 10 月 14 日起到自变量 *year*、*month*、*day* 表示的日期的天数。

## 5. 时间间隔生成函数(属于创建的持续时间函数组)

(1) TIME.DAYS(*days*) 数值型函数。函数值为与自变量 *days* 指定的天数相应的时间间隔。自变量必须是数值型。要将结果显示成时间,就要赋予结果变量时间格式。函数值是与自变量值相应的秒数。例如, TIME.DAYS(3) 的结果是 259200, 当赋予结果变量时间格式 hh:mm:ss 时, 结果为 72:00:00。

(2) TIME.HMS(*hours*) 数值型函数。函数值为与时间间隔变量 *hours* 指定的小时数相应的秒数。*hours* 的值必须是整数; 所有自变量必须处理成或者都是正值, 或者都是负值。要把它显示成时间, 需要赋予结果变量时间格式。在函数列表中该函数名为 TIME.HMS(1)。例如, TIME.HMS(48) 值为 172800, 赋予其时间格式 hh:mm 时, 结果为 48:00。在保持结果变量为数值型时自变量可以是负值, 其他格式将显示为缺失值。

(3) TIME.HMS(*hours, minutes*) 数值型函数。函数值是与时间间隔自变量 *hours*、*minute* 相应的秒数。*hours* 必须是整数; *minutes* 必须是小于 60 的整数。要函数值显示为时间, 则应赋予结果变量时间格式。自变量可以都是负值或都是正值。对于都是负值的自变量, 结果变量只能是数值型, 其他格式将显示为缺失值。在函数列表中该函数名为 TIME.HMS(2)。例如, TIME.HMS(96,30), 结果为 347400, 当赋予结果变量时间格式 hh:mm 时, 显示 96:30。

(4) TIME.HMS(*hours, minute, second*) 数值型函数。函数值是与时间间隔自变量 *hours*、*minute*、*second* 相应的秒数。*hours* 必须是整数; *minutes* 必须是小于 60 的整数; *seconds* 可以包括小数, 但必须小于 60。要函数值显示为时间, 则应赋予结果变量时间格式。自变量可以都是负值或都是正值。对于都是负值的自变量, 结果变量只能是数值型, 其他格式将显示为缺失值。在函数列表中该函数名为 TIME.HMS(3)。例如, TIME.HMS(96, 30, 20.50), 结果为 347420.50, 当赋予结果变量时间格式 hh:mm:ss 时, 显示 96:30:20。

## 6. 时间间隔提取函数(属于时间段提取函数组)

(1) CTIME.DAYS(*timevalue*) 数值型函数。返回给定时间(被看做秒数)值折合的天数(自 1582 年 10 月 14 日算起的天数), 包括分数的天数。自变量 *timevalue* 时间值必须是一个数值或是 SPSS 格式的时间表达式, 如 TIME.×××函数的计算结果。例如, CTIME.DAYS(10800) 值为 0.125 天或 CTIME.DAYS(time.hms(3)), 结果相同。

(2) CTIME.HOURS(*timevalue*) 数值型函数。返回给定时间值折合的带有小数部分的小时数。自变量时间值必须是一个秒数值或 SPSS 格式的时间表达式, 如 TIME.×××函数创建的时间值或用 TIME 输入格式读取的数值。例如, CTIME.HOURS(172830) 结果为 48.008, 显示值的近似程度取决于数值格式的小数位数。

(3) CTIME.MINUTES(*timevalue*) 数值型函数。返回给定时间值折合的带有小数部分的分钟数。自变量时间值必须是一个数值或 SPSS 格式的时间表达式, 例如 TIME.×××函数创建的时间值或用 TIME 输入格式读取的数值。

(4) CTIME.SECONDS(*timevalue*) 数值型函数。返回给定时间值折合的带有小数部分的秒

数。自变量时间值必须是一个数值或 SPSS 格式的时间表达式, 如 `TIME.×××` 函数创建的时间值或用 `TIME` 输入格式读取的数值。

(5) `XDATE.HOUR(datevalue)` 数值型函数。函数值为与自变量 *datetime* 相应的小时数, 一个 0~23 之间的整数。自变量是描述时间或日期时间值。自变量可以是一个数值、时间或日期时间变量或者处理成时间或日期时间值的表达式。

(6) `XDATE.MINUTE(datevalue)` 数值型函数。函数值为表现为时间或日期时间的值相应的 0~59 间的分钟数。自变量可以是一个数值、时间或日期时间变量或者可以是一个已经处理成时间或日期时间值的表达式。

(7) `XDATE.SECOND(datetime)` 数值型函数。函数值为表现为时间或日期时间的值相应的 0~59 间的秒数。自变量可以是一个数值、时间或日期时间变量或者可以是一个已经处理成时间或日期时间值的表达式。

(8) `XDATE.TDAY(timevalue)` 数值函数。函数值是与描述时间或日期时间的数值相应的整天数。自变量可以是一个数值, 时间格式的变量或者处理成时间间隔的表达式。

## 7. 与日期时间有关的转换函数

`NUMBER(stringDate, Date11)` 数值型函数。把内容为标准格式(dd-mmm-yyyy)日期的字符串转换成描述该日期的秒数。如果字符串不能使用标准格式读取, 函数值是系统缺失值。

第一个自变量是字符型, 自变量的值为与 `Date11` 格式相应的日期。

如果定义了字符串格式的自变量, 输入了与 dd-mmm-yyyy 相应的日期, 可以使用该函数将字符串变量转换为日期变量。

## 5.2 日期时间函数的应用

### 5.2.1 日期时间型变量的格式转换

因为日期时间型变量在机内就是从 1582 年 10 月 14 日 24:00:00 算起到达变量值代表的日期时间中的秒数。在计算机内就是一个数值, 只是在显示方式上有所不同。因此, 日期时间型变量与数值型变量的转换只是显示方式的转换, 只要在数据窗的变量窗中改变变量的类型即可。

**【例 1】** 将数据文件 `data05-01.sav` 中的日期型变量 `birthday` 转换为数值型变量, 方法如下:

(1) 图 5-1(a) 所示是数据窗中的日期型变量 `birthday` 显示成 4 位年、2 位月、2 位日的日期格式。

(2) 在【变量视图】窗口建立数值型变量 `birthday1`, 如图 5-1(b) 所示。

(3) 回到【数据视图】窗口, 选择 `birthday` 变量中的所有数据, 在右键菜单中选择【复制】。

(4) 将插入点光标置于 `birthday1` 的第一个观测处, 在右键菜单中选择【粘贴】。在 `birthday1` 中的所有数据均为数值型。

每个观测的 `birthday1` 值均为 `birthday` 值的数值型数值。

也可以先将 `birthday` 的数据全部复制到一个新日期型变量中, 再修改新变量的类型为数值型。

变量 `birthday1` 中的数值是从 1582 年 10 月 14 日 24:00:00 (10 月 15 日 0:00:00) 算起到 `birthday` 变量值指定日期的秒数值。日期型变量到数值型变量转换完毕后的数据文件是 `data05-01a`。

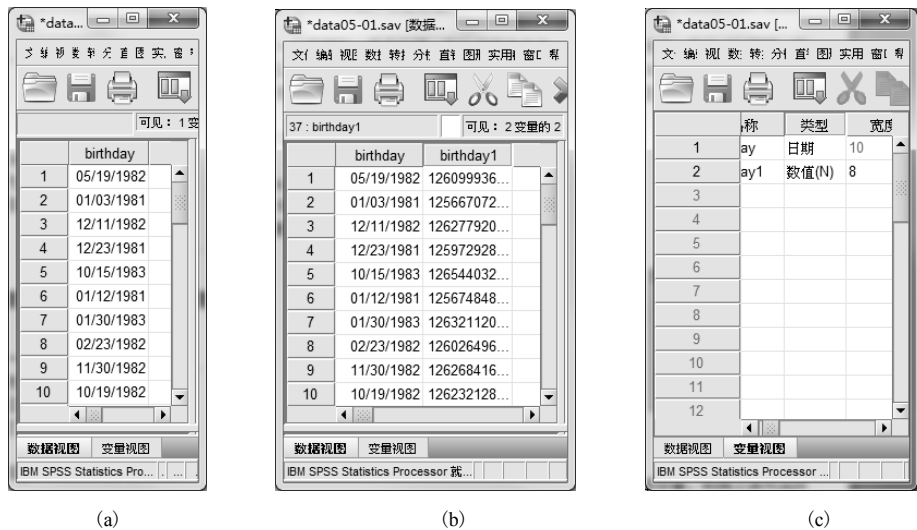


图 5-1 日期型变量转换成数值型变量的过程

【例 2】数据文件 data05-02 中的变量 data1 是数值型变量，如图 5-2(a) 所示。要求将其改变为日期型变量，并在变量类型对话框中指定一种日期时间格式。

步骤如下：

- (1) 单击【变量视图】选项卡，如图 5-2(b) 所示，单击【类型】中的【数值】单元格，打开【变量类型】对话框。如图 5-2(c) 所示。
- (2) 在【变量类型】对话框中选择【日期】型，并在打开的右边的矩形框中选择一种日期时间格式【yyyy/mm/dd】，如图 5-2(c) 所示。

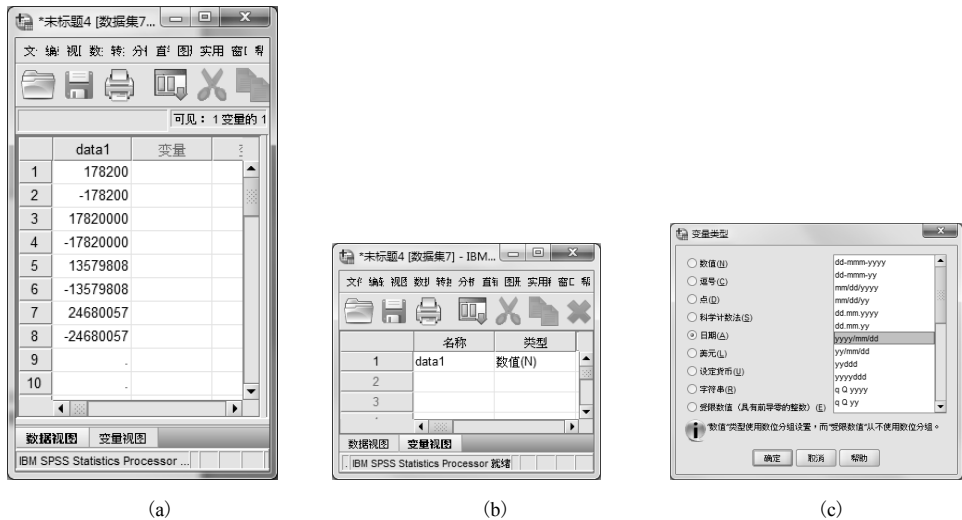


图 5-2 数值型变量转换成日期时间变量的过程

(3) 在【变量类型】对话框中单击【确定】按钮，结果如图 5-3 所示。

为了对比，把转换前后的变量放在一个数据文件中，见图 5-4。负数值转换成日期型，结果是缺失值。转换后的结果见数据文件 data05-02a。





图 5-3 转换后的变量视图和数据视图

【例 3】 字符型变量与日期型变量的转换。

当前日期时间函数产生的是字符型常数。字符型是不能参与算术运算的。因此如果运算涉及由当前日期函数产生的变量，必须先将显示成日期的字符型变量转换成日期型变量。

图 5-5(a)、(b)所示是使用\$Date11 当前日期函数产生的 currentdate 字符型变量，见数据文件 data05-03。



(a)

(b)

图 5-4 数值型转换成日期型变量的结果对比

图 5-5 当前日期时间函数生成的字符型变量

转换为数值型变量的操作步骤是：

- (1) 单击【转换→计算变量】，打开【计算变量】对话框，见图 5-6。
- 将显示为日期形式的字符型变量转换成数值型或日期型变量的方法可利用转换函数来实现。
- (2) 在【计算变量】对话框中：
  - ① 在【目标变量】栏中输入新变量名 currenttime1。
  - ② 单击【类型与标签】按钮，打开【定义新变量类型】对话框。在【标签】栏内输入标签“日期字符转换数值”。【类型】栏内选择【数值】，说明新变量是数值型。
- 单击【继续】按钮，返回【计算变量】主对话框。



图 5-6 【计算变量】对话框

- ③ 在【函数组】栏中选择【转换】的函数。
- ④ 在【函数和特殊变量】栏内选择【Number】，单击向上箭头按钮。显示为【Number(?,?)】，光标停留在第一个问号处。  
在变量列表中选择【currentdate】，单击向右箭头按钮，在第二个问号处输入日期变量格式 Date11。就此形成等式  
`currentdate1=number(currentdate,date11)`  
格式参数也可以是 date9，即  
`currentdate1=number(currentdate,date9)`
- ⑤ 单击【确定】按钮，在数据窗口中生成新变量 currentdate1，转换后的数据见图 5-7，参见数据文件 data05-03a。



图 5-7 字符型日期值转换成数值型变量的结果

5.2.2 日期时间型变量的算术运算

- 【例 4】 业余体校某项运动的校友花名册中记录了老运动员的出生日期。计算到当前日期为止，这些老队员的年龄。步骤如下：
- (1) 在【变量视图】窗口中建立一个变量名为 birthday 的变量。

(2) 在【类型】栏内定义该变量为日期型。选择下拉菜单中的【日期】，打开【选择格式】对话框，在对话框中选择【yyyy/mm/dd】，单击【确定】按钮返回【变量视图】窗口，自动显示变量宽度 10，见图 5-8(a)。

(3) 在【数据视图】窗口中，按选择的 yyyy/mm/dd 格式输入运动员生日，如图 5-8(b)所示，见数据文件 data05-04。

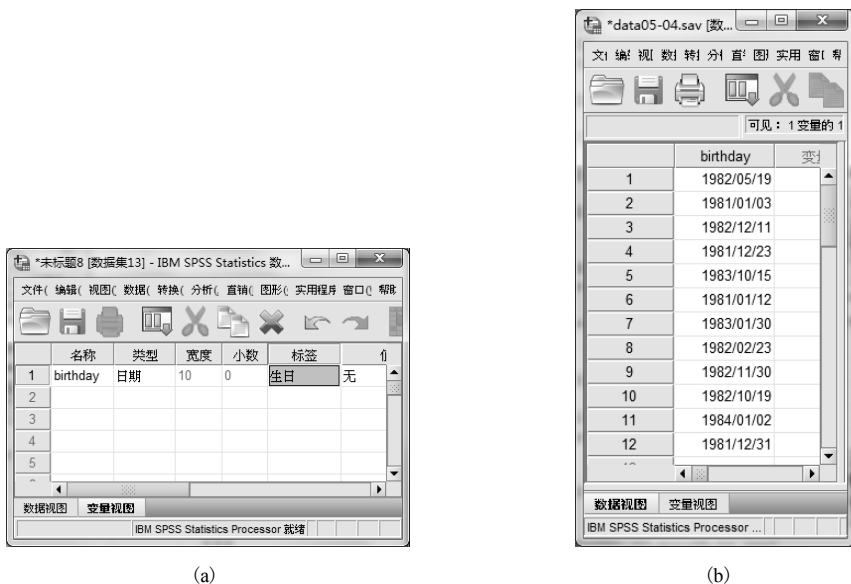


图 5-8 定义一个日期变量，输入日期数据

(4) 单击【转换→计算变量】打开【计算变量】对话框；在【目标变量】栏内输入新变量名“Curda”；单击【类型与标签】按钮，打开【定义变量类型与标签】对话框。在【标签】栏输入变量标签“当前日期”；在【类型】栏内选择【字符串】，表明新变量是字符型。因为在该对话框内只能在数值型和字符型中选择一个，而由于日期的表达中分隔符的存在，不可能是数值型。宽度输入“10”，单击【继续】按钮。

(5) 按上一节所述的方法将字符型的 Curda 当前日期变量转换为数值型，变量名为【currentdate】，表达式为【NUMBER(curda,date11)】，如图 5-6 所示，见数据文件 data05-04a。

也可以一步完成以上操作：currentdate= NUMBER(\$DATE11,date11)，无须再建立字符串变量 curta。

(6) 由于变量 birthday 是日期型变量，而它的机内置是数值，currentdate 是数值型变量都是自 1582 年 10 月 14 日 24:00:00 算起的秒数，所以可以进行算术运算。

(7) 调用日期计算函数，得到年龄变量 age。

单击【转换→计算变量】，打开相应的对话框。

① 在【目标变量】栏内输入新变量名“age”；单击【类型与标签】按钮，打开【计算变量：类型与标签】对话框。在【标签】栏输入变量标签“年龄”，在【类型】栏选择数值，单击【继续】按钮返回主对话框。

② 在【计算变量】对话框中，在【函数组】栏中选择【日期运算组】，在【函数和特殊变量】栏内选择计算日期差函数【Datediff】，单击向上箭头按钮，该函数显示在【数字表达式】栏中，即【Datediff(?,?,?)】。

③ 在【原始变量】栏内选择当前日期 Currentdate 变量作为第一个参数，单击向右箭头按钮，代替函数中第一个问号；同样方法选择 birthday 作为第二个参数变量，代替第二个问号；在第三个问号处输入“years”作为第三个参数。注意，必须带半角引号。在【数字表达式】栏内显示调用函数【DATEDIFF(Currentdate,birthday,"years")】。单击【确定】按钮。

④ 在【计算变量】窗中察看新变量 age 的结果，如图 5-9 所示，见数据文件 data05-04a。



图 5-9 【计算变量】对话框中生成当前日期变量

以上操作，可以使用函数嵌套方式组成一个表达式：

Age= datediff(number(\$date11,date11),birthday,"years")

在【计算变量】对话框中，先定义一个新变量 age，再在【数字表达式】框中输入以上表达式或者通过选择函数并设定函数参数的方法实现上述嵌套函数的输入，即可一步获得 age 变量的值，见数据文件 data05-04b。

**【例 5】** 班委会决定在每个月为在该月份过生日的同学时举办一次庆祝活动，班级花名册中记载着每个同学的生日，在数据窗中对应变量 birthday。为了统计每个月有几个人过生日，需要把生日的月份提取出来。应该如何操作？

见数据文件 data05-05。

首先确定需要使用的是提取月份的函数 XDATE.MONTH(datevalue)，操作步骤如下：

(1) 单击【转换→计算变量】，打开【计算变量】对话框。

(2) 在【目标变量】栏内输入新变量名“birthday1”；单击【类型与标签】按钮，打开【计算变量类型与标签】对话框。在【标签】栏输入变量标签“生日月份”，在【类型】栏选择数值，单击【继续】按钮返回【计算变量】主对话框。

(3) 在【函数组】栏中选择【抽取日期类】，在【函数与特殊变量】栏内选择提取月份的函数【XDATE.MONTH】，单击向上箭头按钮，该函数显示在【数字表达式】栏中，即【XDATE.MONTH(?)】。

(4) 在【原始变量】栏内选择【birthday】作为函数自变量，单击向右箭头按钮，代替函数中的问号，显示为【XDATE.MONTH(birthday)】。

(5) 单击【确定】按钮，在数据窗中显示变量 birthday1 的值，即每个同学生日中的月份，见图 5-10，结果保存在数据文件 data05-05a 中。



	birthday	month	变量
1	05/19/1982	5	
2	01/03/1981	1	
3	12/11/1982	12	
4	12/23/1981	12	
5	10/15/1983	10	
6	01/12/1981	1	
7	01/30/1983	1	
8	02/23/1982	2	
9	11/30/1982	11	
10	10/19/1982	10	

图 5-10 提取月份的结果

## 习 题 5

1. 上网查询，为什么 SPSS 的时间运算总是与 1582 年 10 月 14 日 24:00:00 (10 月 15 日 0:00:00) 有关，即日期型数据转换成以这个时间点为起始点的秒数？
2. 定义一个日期变量，输入你们班同学的生日，计算他们此时的年龄。
3. 计算你们班同学中生日在 10 月份的人数。

# 第6章 构建表格

## 6.1 自定义表格

### 6.1.1 自定义表格的概念

表格是数据资料整理中一种很常用的形式。一个好的表格能使统计资料系统化、层次化和条理化。虽然在运行 SPSS 的各种分析过程中，也能在输出窗口自动产生各种表格，但表格的形式不一定能满足撰写报告的要求，因此，在很多场合，会根据不同表述的需要，用到自定义的各种表格。

在 SPSS 20.0 中，在其【分析】下拉菜单中的【(制)表】命令的【设定表】过程(见图 6-1)中，可以通过自定义表格的方式产生 1、2、3 维表格。各维度可用单个变量或变量组合来定义。在每一维(行、列和层)中，可以叠放多重变量使之成为复合表，并可为嵌套变量建立子表。

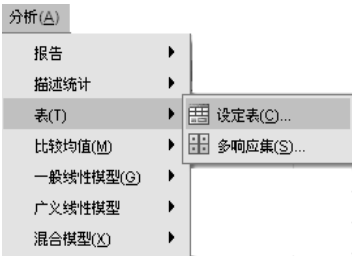


图 6-1 自定义表格过程

#### 1. 表格的组成

常用的二维表格通常由行和列交叉组成。根据表格行、列中变量嵌套的情况，可将表格分成简单和复杂两种。最简单的表格形式可由单列两行或单行两列组成。表 6-1 是一张较复杂的表格，它由男、女对婚姻幸福程度感受的同结构的两个单表上下叠加而成。

表 6-1 不同性别多子女人群对婚姻幸福程度的感受

				各组分类中子女的数量			
				0	1~2	3~4	>=5
				列%	列%	列%	列%
性别	男	婚姻幸福程度	很幸福	36.2	64.9	62.2	71.4
			中等幸福	32.9	33.0	35.2	26.2
			不太幸福	3.9	2.1	2.4	2.4
	女	婚姻幸福程度	很幸福	69.2	63.9	61.6	57.6
			中等幸福	28.6	33.2	35.5	23.3
			不太幸福	2.2	2.9	2.8	9.1

表格通常包括以下几个部分：

- (1) 表头。一般位于表格的上方，简明地描述表格所反映的中心内容。
- (2) 行变量。定义表格的行，在表格左边由上到下排列的变量称为行变量。
- (3) 列变量。定义表格的列，在表格上面横向排列的变量称为列变量。
- (4) 单元格。由表格的行和列的交叉点形成。

(5) 表格的实体部分。由所有单元格组成，包括由表格总计、总和、平均数、百分比等基本信息。

(6) 角注位于右上角，脚注位于表格的下方。简要说明表格的组成、生成的日期、时间或其他需要特别的声明。

2. 表格的结构类型

根据表格中变量的位置、作用可将表格分为以下几种类型：

- (1) 简单表格。由一个变量和若干统计汇总指标组成，变量可以在行或在列上。
- (2) 简单交叉表。在行、列上均设置了一个变量，包括各变量的分类、统计汇总指标。
- (3) 堆栈表格。在行(或列)上有并列的两个以上变量，见图 6-2(a)。
- (4) 嵌套表格。同行或同列上由不同层次上的变量组成的表格，见图 6-2(b)。
- (5) 分层表格。按某变量的分类分别形成表格，每层一个表，见图 6-2(c)。

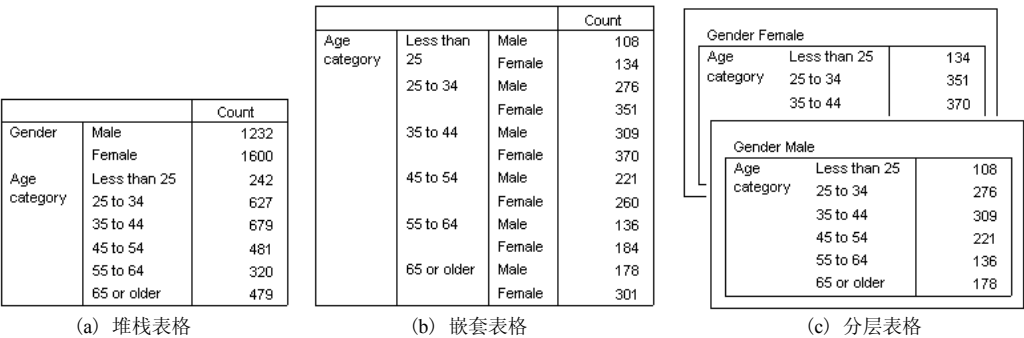


图 6-2 表格结构

3. 变量的测度方法与设定表程序能自动识别的测度标准

在 SPSS 中，测度方法共有 3 种，即序号测度(习惯称有序测度)、名义测度和度量测度(习惯称尺度测度)。用前两种方法测度的变量，也即有序变量和名义变量是分类变量，它们可以定义表格的行、列和层。默认的汇总统计指标是计数；尺度变量一般被汇总在分类变量的类别里，默认的汇总统计指标是均值(算术平均数)。在没有使用分类变量定义组别时，还可以用尺度变量本身来汇总尺度变量。它主要用于多重尺度变量的分层汇总。

6.1.2 自定义表格的操作

1. 自定义表格的操作

按【分析→表→设定表】顺序单击菜单项，第一次启动时，首先打开的是图 6-3 所示的【设定表格】预备对话框。

在进入制表工作之前，一般应先定义好变量的各种属性，尤其是变量类型、变量标签、值标签、测定标准等。如果这一步工作没做好，则可在【设定表格】预备对话框中，单击【定义变量属性】按钮，在其后的对话窗口中将变量属性定义完整。相关内容已在第 2 章介绍过，具体做法请参阅 2.1.6 节。如果变量属性已经定义，则可以选择【不再显示此对话框】选项并单击【确定】按钮关闭这一对话框，进入【设定表格】主对话框，见图 6-4。有 4 个选项卡，分别是【表格】、【标题】、【检验统计量】和【选项】。

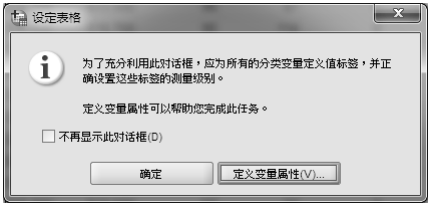


图 6-3 【设定表格】预备对话框



图 6-4 【设定表格】主对话框

2. 简单的制表操作过程

表格在【表格】选项卡中构建。从【变量】列表中拖曳一个变量或多应答变量集到右侧的探究窗口行或列的区域，就可以建立一个最简单的表格。探究窗口显示将要建立的表格的轮廓，供制表人预览，不显示实际数据值。

如果拖曳到行或列中的是分类变量，则在探究窗口中显示该变量的各分类值。与此同时，在探究窗口中显示默认的统计量名称：计数。

如果拖曳到行或列中的是尺度变量，则只在探究窗口的行(或列)中显示变量名和默认的统计量名称：均值。单击【确定】按钮，所建表格显示在输出窗口中。

3. 各种结构的表格

(1) 如要选择多个变量，并且将它们一起拖曳置于探究窗口的行(或列)中，则所选择变量各分类在行(或列)中是并排的，生成堆栈表格。

(2) 如果将变量拖到已在窗口中的行变量左面(或右面)，形成行嵌套，左面变量在外层。如果将变量拖至窗口中列变量的上方(或下方)，形成列嵌套，上面的变量在外层。

(3) 表格的分层。当拖曳一个变量到探究窗口右上方的【层】图标处时，该变量被定义为层变量。无论测度方法是有序的还是标称的，分类变量都可以作为层变量。层变量的每个分类的数据按探究窗口中设置的行、列变量构成一个表格，称为一层，每层结构相同。层变量数及其分类数决定表格数量，要慎重指定层变量。层变量之间是并列关系，层数等于各层变量分类数之和，例如 3 个 3 类分层变量共 9 层；层变量之间是嵌套关系，层数是各变量分类数之积，例如 3 个 3 类变量生成 27 层，即 27 个表格。由此可知层变量及每个层变量的分类数与生成表格数之间的关系，故层变量的选择需慎重。

(4) 从【探究】窗口中选择一个变量，按 Delete 键或将其拖曳到窗口外即从表格中删除该变量。

6.2 汇总、统计指标与统计检验

6.2.1 统计指标与汇总项

在【表格】选项卡中，每定义一个行变量或列变量，就可以马上定义要在表格中出现的该变量的统计指标和汇总项。



1. 定义栏

在【定义】栏中有两个选项，在【探究】窗口选择行、列上的尺度变量时，激活【摘要统计量】选项；选择分类变量时，激活【分类和总计】选项。

(1) 摘要统计量

在【探究】窗口选择尺度变量，单击【摘要统计量】按钮，进入定义这个变量的统计量的对话框，见图 6-5。在该对话框中，每选择一个统计量，单击【箭头】按钮送入【显示】表中，便可以在【显示】表中编辑这个统计量的标签 (SPSS 汉化为“标链”)，从【格式】下拉列表选取统计指标的显示格式，输入小数位数。单击右边的上下箭头按钮改变当前统计指标的显示顺序。



图 6-5 【摘要统计】对话框

单击【应用选择】按钮，在生成的表格中只对当前选择的变量计算指定的汇总统计指标；单击【应用到全部】按钮，对探究窗口所有同类型的变量都计算指定的汇总统计指标。

汇总统计指标的有效性取决于所选择变量的测定标准和变量的选择顺序。对嵌套表格来说，对嵌套在最内层的变量有效。分层表格对每层中的指定变量都有效。

① 适合所有变量的汇总统计指标见表 6-2。

表 6-2 适合所有变量的汇总统计指标

指 标	计算内容描述	默认标签	默认格式
COUNT <sup>①</sup>	各类中样品的数量。它对分类和多重应答变量是默认的	计数	nnnn
ROWPCT.COUNT	基于单元格计数的行百分比。在子表格内部计算	行 N%	nnnn.n%
COLPCT.COUNT	基于单元格计数的列百分比。在子表格内部计算	列 N%	nnnn.n%
TABLEPCT.COUNT	基于单元格计数的表格百分比	表 N%	nnnn.n%
SUBTABLEPCT.COUNT	基于单元格计数的子表格百分比	子表 N%	nnnn.n%
LAYERPCT.COUNT	基于单元格计数的层百分比。未定义层时同表格百分比	层 N%	nnnn.n%
LAYERROWPCT.COUNT	基于单元格计数的行百分比。整行(即子表)的百分比总和为 100%	层行 N%	nnnn.n%
LAYERCOLPCT.COUNT	基于单元格计数的列百分比。整列(即子表)的百分比总和为 100%	层列 N%	nnnn.n%

续表

指 标	计算内容描述	默认标签	默认格式
ROWPCT.VALIDN	基于有效计数的行百分比	行有效 N%	nnnn.n%
COLPCT.VALIDN	基于有效计数的列百分比	列有效 N%	nnnn.n%
TABLEPCT.VALIDN	基于有效计数的表格百分比	表有效 N%	nnnn.n%
SUBTABLEPCT.VALIDN	基于有效计数的子表格百分比	子表有效 N%	nnnn.n%
LAYERPCT.VALIDN	基于有效计数的层百分比	层有效 N%	nnnn.n%
LAYERROWPCT.VALIDN	基于有效计数的行百分比。整行的百分比和为 100%	分层行有效 N%	nnnn.n%
LAYERCOLPCT.VALIDN	基于有效计数的列百分比。整列的百分比和为 100%	分层列有效 N%	nnnn.n%
ROWPCT.TOTALN	基于总计数的行百分比，包括读者和系统缺失值	行总计 N%	nnnn.n%
COLPCT.TOTALN	基于总计数的列百分比，包括读者和系统缺失值	列总计 N%	nnnn.n%
TABLEPCT.TOTALN	基于总计数的表格百分比，包括读者和系统缺失值	表格总计 N%	nnnn.n%
SUBTABLEPCT.TOTALN	基于总计数的子表格百分比，包括读者和系统缺失值	子表总计 N%	nnnn.n%
LAYERPCT.TOTALN	基于总计数的层百分比，包括读者和系统缺失值	层总计 N%	nnnn.n%
LAYERROWPCT.TOTALN	基于总计数的行百分比，包括读者和系统缺失值。整行的百分比和为 100%	层行总计 N%	nnnn.n%
LAYERCOLPCT.TOTALN	基于总计数的列百分比，包括读者和系统缺失值。整列的百分比和为 100%	层列总计 N%	nnnn.n%

注：① 在美国英语体系中这是默认的。后缀.COUNT 可以从计算基于单元格的百分比中省略。因而 ROWPCT 等于 ROWPCT.COUNT。

② 适合于分类变量的汇总统计指标：

- 计数，各单元格中样品的数量，或多重应答集中应答的数量。
- 未加权的计数，表格的每个单元格中样品的未加权的数量。
- 列 N%，子表每列的百分数之和为 100%。仅当有分类行变量时，列百分数才是有效的。
- 行 N%，子表每行的百分数之和为 100%。仅当有分类列变量时，行百分数才是有效的。
- 层行和层列 N%，每层中行百分比和列的百分比。嵌套表格中各子表的行或列百分数的总和为 100%；分层表格每层中所有嵌套子表的行或列的百分数总和为 100%。
- 层 N%，每层内的百分数。作为简单百分数，当前可见层里单元格的百分数总和为 100%。如果没有层变量，则它等于同表格百分数。
- 表 N%，表中各单元格的百分数建立在整个表格基础上。所有单元格的百分数是建立在样品总数的基础上的，并且整个表格百分数的总和为 100%。
- 子表 N%，子表中每个单元格的百分数建立在子表基础上。子表中，所有单元格的百分数建立在子表内相同样品总数的基础上，并且在子表内单元格的百分数总和为 100%。在嵌套表中，在最里面嵌套水平之前的变量定义子表格。百分数受计算它们的基数(分母)的影响，并且选项的数量决定基数。基于样品、应答或计数，多重应答集可以有百分数。

由层变量定义的各层表格被当做独立的表格处理。在各层表格内，各层的行 N%的总和、各层的列 N%的总和以及每层的表 N%的总和都为 100%。

③ 适合尺度变量及分类自定义合计的主要统计指标见表 6-3。对分类变量还能求部分和及自定义求总和。表格默认包括总计或小计。

④ 适合多重应答集的统计指标见表 6-4。

表 6-3 适合尺度变量、合计和小计的主要统计指标

指 标	描 述	默认标签	默认格式
MAXIMUM	最大值	最大值	自动
MEAN	算术平均数。默认尺度变量	均值	自动
MEDIAN	中位数	中位数	自动
MINIMUM	最小值	最小值	自动
MISSING	缺失值合计(用户和系统的缺失值)	缺失	自动
MODE	众数。如果有结(众数相同)，则显示最小的值	众数	自动
PTILE	百分位数。取 0~100 间的一个数值作为需要的参数。 PTILE 在 SPSS Tables 中是用同样的 PTILE 来计算的。注意， 在 SPSS Tables 中默认的百分位数的方法是 HPTILE	百分位数 nn	自动
RANGE	两极差	范围 (应汉化为两极差)	自动
SEMEAN	标准误	标准平均误差 (应汉化为标准误)	自动
STDDEV	标准差	标准差	自动
SUM	数值总和	合计	自动
TOTALN	非缺失值、用户缺失值和系统缺失值的合计。由 CATEGORIES 子命令中隐含的有效值不计数	总计 N	nnnn
VALIDN	非缺失值合计	有效 N	nnnn
VARIANCE	方差	方差	自动
ROWPCT.SUM	基于总和的行百分比	行和%	nnnn.n%
COLPCT.SUM	基于总和的列百分比	列和%	nnnn.n%
TABLEPCT.SUM	基于总和的表格百分比	表和%	nnnn.n%
SUBTABLEPCT.SUM	基于总和的子表格百分比	子表合计%	nnnn.n%
LAYERPCT.SUM	基于总和的层百分比。	层和%	nnnn.n%
LAYERROWPCT.SUM	基于总和的行百分比。整行的百分比总和为 100%	分层行合计%	nnnn.n%
LAYERCOLPCT.SUM	基于总和的列百分比。整列的百分比总和为 100%	分层列合计%	nnnn.n%

表 6-4 适合多重应答集的主要统计指标

指 标	计算内容描述	默认标签	默认格式
RESPONSES	应答合计	响应	nnnn
ROWPCT. RESPONSES	行百分比，回答合计是分子。回答的总计是分母	行响应%	nnnn.n%
COLPCT. RESPONSES	列百分比，回答合计是分子。回答的总计是分母	列响应%	nnnn.n%
TABLEPCT. RESPONSES	表格百分比，回答合计是分子。回答的总计是分母	表格响应%	nnnn.n%
SUBTABLEPCT. RESPONSES	子表格百分比，回答合计是分子。回答的总计是分母	子表响应%	nnnn.n%
LAYERPCT. RESPONSES	层百分比，回答合计是分子。回答的总计是分母	层响应%	nnnn.n%
LAYERROWPCT. RESPONSES	层中行百分比，回答合计是分子。回答的总计是分母，子表格中整行的百分比和为 100%	分层行响应 %	nnnn.n%
LAYERCOLPCT. RESPONSES	层中列百分比，回答合计是分子。回答的总计是分母，子表格中整列的百分比和为 100%	分层列响应%	nnnn.n%
ROWPCT.RESPONSES. COUNT	行百分比：行的回答合计是分子，总计为分母	行响应% (基于：计数)	nnnn.n%
COLPCT.RESPONSES. COUNT	列百分比：列的回答合计是分子，总计为分母	列响应% (基于：计数)	nnnn.n%
TABLEPCT. RESPONSES.COUNT	表格百分比：表格的回答合计是分子，总计为分母	表格响应% (基于：计数)	nnnn.n%
RESPONSES	分母	(基于：响应)	
SUBTABLEPCT. COUNT.RESPONSES	子表百分比：子表中的合计是分子，回答的总计是分母	子表计数% (基于：响应)	nnnn.n%

续表

指 标	计算内容描述	默认标签	默认格式
LAYERPCT. COUNT.RESPONSES	层百分比：层中的合计是分子，回答的总计是分母	层计数% (基于：响应)	nnnn.n%
LAYERROWPCT. COUNT.RESPONSES	行百分比：行中的合计是分子，回答的总计是分母。 整行(即子表格)的百分比和为 100%	分层行计数% (基于：响应)	nnnn.n%
LAYERCOLPCT. COUNT.RESPONSES	列百分比：列中的合计是分子，回答的总计是分母。 整列(即子表格)的百分比和为 100%	分层列计数% (基于：响应)	nnnn.n%
SUBTABLEPCT. RESPONSES.COUNT	子表格百分比：子表的回答合计是分子，总计为分母	子表响应% (基于：计数)	nnnn.n%
LAYERPCT. RESPONSES.COUNT	计算层百分比：层的回答合计是分子，总计为分母	层响应% (基于：计数)	nnnn.n%
LAYERROWPCT. RESPONSES.COUNT	计算行百分比：行的回答合计是分子，总计为分母。 整行(即子表格)的百分比和为 100%	分层行响应% (基于：计数)	nnnn.n%
LAYERCOLPCT. RESPONSES.COUNT	计算列百分比：列的回答合计是分子，总计为分母。 整行(即子表格)的百分比和为 100%	分层列响应% (基于：计数)	nnnn.n%
ROWPCT.COUNT. RESPONSES	行百分比：行的回答合计是分子，回答的总计是分母	行计数% (基于：响应)	nnnn.n%
COLPCT.COUNT. RESPONSES	列百分比：列中的回答合计是分子，回答的总计是分母	列计数% (基于：响应)	nnnn.n%

⑤ 设定关于总计和小计的摘要统计量。使用【摘要统计量】对话框，在表格的【总计】和【小计】区域，可以显示除总计之外的其他统计指标。

(2) 分类与总计

当拖曳一个分类变量或多重应答集到探究窗口后，单击【定义】栏中的【分类和总计】按钮，进入相应的对话框，见图 6-6。

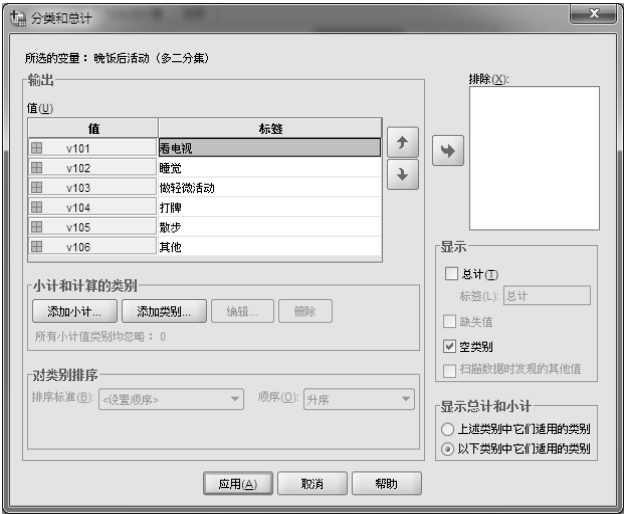


图 6-6 【分类和总计】对话框

① 【输出】和【排除】栏。在【输出】栏中显示生成的表格中分类变量的所有值和值标签。单击上、下箭头，可移动分类值在表中的位置。若要生成表格不包括某个分类值，将其移入【排除】栏中。

② 【对类别排序】。在该栏中，【排序标准】后的下拉列表中可选择排序依据，可以选择按

【值】、【标签】或【计数】(单元格频数)排序;在【顺序】下拉列表中选择决定排序是【升序】还是【降序】。

③【小计和计算的类别】。如果需要对部分分类值计算小计,只需在【输出】栏中选择一个分类值,单击【添加小计】按钮,在弹出的【定义小计】对话框(见图 6-7)中,对标题可以重新命名,按【继续】按钮,则小计项插入到【输出】栏中所选分类值的下方,并在【值】列中显示小计的范围。同样,单击【添加类别】按钮,则弹出【定义计算的类别】对话框,见图 6-8。可在【已计算类别】【标签】框中定义小计的标签,将所要自定义的计算类别用表达式的方式,如[1]+[3]输入到【已计算类别的表达式】框中,表示要计算第 1 类和第 3 类的小计。单击【继续】按钮,则返回图 6-6 所示的【分类和总计】对话框,小计项插入【输出】栏中所选分类值的下方;单击【编辑】按钮,则返回图 6-8 所示的【定义计算的类别】对话框,可对前面在此对话框中作过的设置进行回顾和修改;如果要移走添加的类别,单击【删除】按钮即可。

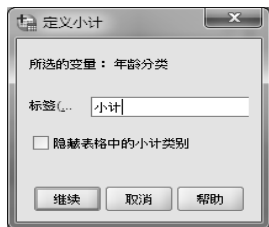


图 6-7 【定义小计】对话框



图 6-8 【定义计算的类别】对话框

④【显示】。【空类别】即频数为 0 的分类、【扫描数据时发现的其他值】这两个选项为系统默认选项。如果选择【显示】栏中的【总计】,则可以继续选择特殊分类总计的内容。【缺失值】选项不可用。

⑤【显示总计和小计】。用这里的选项可以重新定义③中所添加的小计项的统计范围。相对于被小计(或被总计)的所有分类值的位置,选择【上述类别中它们适用的类别】,则对③中所添加的小计项所在位置下方的所有适用类别进行小计(或总计);选择【以下类别中它们适用的类别】,则对③中所添加的小计项所在位置上方的所有适用类别进行小计(或总计)。即用这两个选项来说明所插入的小计项是在所要小计的分类值的上方还是下方。软件中汉化的两个选项名称不确切。

设置完成单击【应用】按钮,确定选择,返回【设定表格】对话框。

⑥ 在【设定表格】对话框的【摘要统计量】栏中选择统计指标出现的位置和维度。

- 【位置】下拉列表中,有两个选项:选择【列】,统计指标出现在列中;选择【行】,统计指标出现在行中。如果在表中不想出现统计指标,选择【隐藏】选项。
- 【源】下拉列表中,可通过选择【列变量】、【行变量】、【层变量】来改变汇总统计的维度。

- 在【类别位置】下拉列表中，选择【默认】(Default) (软件汉化为“缺失值”不妥)按系统默认格式显示；选择【列中的行标签】或【行中的列标签】则分别为行分类标签显示在列中或列分类值标签显示在行中。
- ⑦ 界面方式的切换按钮在【探究】窗口上方，【普通】界面显示包括在表格中的所有行、列及分类变量的类别和汇总统计等内容；【压缩】界面只显示表格中的变量名及其位置。

2. 层

单击探究窗口右上角的【层】按钮，显示层变量列表。多重应答集被当做分类变量列出。如果有两个以上的层变量，可以在下面两个选项中，确定层的构建方式：

- 【将每个类别显示为一层】，层数为分类数总和。
- 【将每个类别组合显示为一层】，层数为各层变量分类数的乘积。

6.2.2 表格中的统计检验

在主对话框中单击【检验统计量】选项卡，进入如图 6-9 所示的对话框，可以对自定义表格中的变量作不同的显著性统计检验。选项包括：

(1) 【比较列的平均值(t-检验)】。列的均数相等检验。在列中至少有一个分类变量和在行中至少有一个尺度变量的表格可以选择此项。可以选择是否用邦弗伦尼(Bonferroni)方法校正检验的  $P$  值。还可指定检验的【 $\alpha$  水平】，该值应大于 0 且小于 1，系统默认值为 0.05。另外，也可以为多重响应变量所比较的类别估算方差。

(2) 【比较列的比例(z-检验)】。列比率相等检验。行和列中至少存在一个分类变量的表格可以选择此项。可以选择是否用邦弗伦尼方法校正检验的  $P$  值。还可以指定检验的【 $\alpha$  水平】，该值应该大于 0 且小于 1，系统默认值为 0.05。

在上述两种检验过程中，【标识显著性差异】的方式有两种：一种是在单独表中，另一种是在使用 APA 样式下标的主表中。

(3) 【独立性检验(卡方验证)】。独立性卡方检验。行和列中最少有一个分类变量的表格可以选择此项。还可以指定检验的【 $\alpha$  水平】，系统默认值为 0.05。

系统默认检验中包含多重响应变量，也可以选用小计来代替小计类别。



图 6-9 【检验统计量】选项卡

6.3 标题与其他选项

6.3.1 定义表格标题

在【表格自定义】对话框中，单击【标题】选项卡，见图 6-10。

- (1) 在【标题】框中输入表格的标题。

- (2) 在【题注】框中输入脚注。
- (3) 在【角注】框中输入显示在表格左上角的说明文字。只有当定义行变量且当基准行的维度标签已设置成【嵌套】时才显示。这不是默认表格外形的设置。
- (4) 插入当前日期、时间的方法是将插入点光标移至【标题】、【题注】栏中，单击选项卡上方的【日期】或【时间】按钮。
- (5) 【表格表达式】。在【标题】对话框中插入此项标识，产生的表格在相应位置显示各变量在表格中的作用：“+”表示分层变量；“>”表示嵌套变量；“BY”表示交互变量或层变量。



图 6-10 【标题】选项卡

6.3.2 定义表格选项

在【自定义表格】主对话框中单击【选项】选项卡，见图 6-11。



图 6-11 【选项】选项卡

(1) **【数据单元格外观】** 栏。定义空单元及无法进行统计计算的单元里显示什么。

① **【空单元格】** 选项。对计数为 0 的单元能选择显示：**【零(0)】**、**【空】** 或 **【文本】**。说明文本的最大长度为 255 个字符。

② **【不能计算的统计量】**。对指定的统计量不能计算时，如分类里没有样品的均数，相应位置要显示的文字，字符数小于等于 255。默认值用圆点表示。

(2) **【数据列宽度】** 栏，定义数据列最小和(或)最大宽度。

① **【表格外观设置】**。使用默认的表格外观参数中的列宽度。

② **【设定】**。指定最小和最大的列宽度和所使用的单位：**【磅】**、**【英寸】** 或 **【厘米】**。

(3) **【尺度变量缺失值】** 栏。对有两个或更多尺度变量的表格定义计算尺度变量的统计指标时有关缺失数据的处理方法。它有两个选项：

① **【最大限度地使用可用数据(逐个变量删除)】** 选项。在默认表格中主要统计指标的计算包含每个尺度变量所有具有有效值的样品。

② **【在各个尺度变量上使用一致的样品库(剔除法)】** 选项。对表格里的任意尺度变量，计算主要统计指标剔除所有带有缺失值的样品。(注：在 SPSS 中将其汉化为 **【对于尺度变量(整个列表删除)使用一致的个案基础】**。)

(4) **【计算多个类别集中的重复响应】**。对多重应答集中的变量，在默认情况下，不计算重复应答的数量；选择此项，则计算重复应答的数量。

(5) **【隐藏较小计数】**。在该选项下的框中可以输入一个正整数，来定义隐藏的计数。系统默认值为 5。

## 6.4 自定义表格实例

**【例 1】** 对不同性别、年龄、婚姻状况的生活方式和首选早餐的调查研究，数据文件为 data06-01。操作方法如下：

(1) 打开数据文件 data06-01，按**【分析→表→设定表】**顺序单击菜单项，进入**【设定表格】**主对话框。

(2) 将**【变量性别】**(gender)拖曳到**【行】**中，将**【年龄段】**(agecat)拖曳并嵌套在**【性别】**下，将**【婚姻状况】**(marital)拖曳并嵌套在**【年龄段】**下，将**【生活方式】**(Lifestyle)、**【首选早餐】**(bfast)移到**【列】**中。

(3) 单击**【标题】**按钮，进入**【标题】**选项卡。在**【标题】**下面的编辑框中输入表头“不同性别、年龄、婚姻状况的生活方式和首选早餐的统计表”；将插入键移到**【题注】**下面的编辑框中，单击**【日期】**和**【时间】**按钮。设置**【角注】**为制表时的计算机系统的当前日期和时间。

(4) 单击**【检验统计量】**按钮，进入相应的对话框。选择**【独立性检验(卡方验证)】**项，不选择默认的**【隐藏较小计数】**选项。对表格的列变量的分类间作独立性的卡方检验。

(5) 单击**【确定】**按钮，运行程序，在输出窗口中得到如表 6-5 和表 6-6 所示的输出结果。

表 6-5 反映了不同性别、年龄、婚姻状况的生活方式和首选早餐的人数统计。

表 6-6 是皮尔逊卡方检验表。表中列出了不同性别、年龄、婚姻状况下的生活方式和首选早餐两个变量的各项间的皮尔逊卡方独立性检验，两个数据列中第一个数据是卡方值；第二个数据是自由度；第三个数据是在原假设为真的前提下，出现目前统计量值或更加极端值



的概率，换言之，它是拒绝原假设去接受备选假设时，准备犯错误的概率，该值小于 0.05 说明在该性别、年龄、婚姻状况下的生活方式或首选早餐的各项间有差异，其余类推。需要注意的是，本例中，由于 20% 多的期望频数出现小于 5 的情形，因此，独立性检验时最好改用精确检验法。

表 6-5 频数分布表

不同性别、年龄、婚姻状况的生活方式和首选早餐的统计表

						生活方式		首选早餐		
						不活动	活动	早餐吧	麦片	谷类
						计数	计数	计数	计数	计数
性别	男	年龄分类 < 31	婚姻状况	未婚	18	26	25	0	19	
				已婚	14	27	15	0	26	
		31-45	婚姻状况	未婚	10	14	13	2	9	
				已婚	33	40	26	12	35	
		46-60	婚姻状况	未婚	5	13	5	7	6	
				已婚	60	35	16	37	42	
		> 60	婚姻状况	未婚	31	10	4	27	10	
				已婚	65	23	0	70	18	
	女	年龄分类 < 31	婚姻状况	未婚	15	33	27	1	20	
				已婚	23	25	17	3	28	
		31-45	婚姻状况	未婚	12	16	17	1	10	
				已婚	34	47	34	9	38	
		46-60	婚姻状况	未婚	19	5	7	10	7	
				已婚	55	39	11	43	40	
		> 60	婚姻状况	未婚	49	27	10	47	19	
				已婚	31	26	4	41	12	

2013/8/18 17:14:08

表 6-6 皮尔逊卡方检验

Pearson 卡方检验

						生活方式	首选早餐
性别	男	年龄分类 < 31	婚姻状况	卡方	.414	3.487	
				df	1	1	
				Sig.	.520	.062 <sup>b</sup>	
		31-45	婚姻状况	卡方	.092	2.802	
				df	1	2	
				Sig.	.762	.246	
		46-60	婚姻状况	卡方	7.752	1.395	
				df	1	2	
				Sig.	.005 <sup>a</sup>	.498	
		> 60	婚姻状况	卡方	.045	9.482	
				df	1	2	
				Sig.	.832	.009 <sup>a, b</sup>	
	女	年龄分类 < 31	婚姻状况	卡方	2.788	4.606	
				df	1	2	
				Sig.	.095	.100 <sup>b</sup>	
		31-45	婚姻状况	卡方	.007	3.443	
				df	1	2	
				Sig.	.935	.179	
		46-60	婚姻状况	卡方	3.488	4.754	
				df	1	2	
				Sig.	.062	.093	
		> 60	婚姻状况	卡方	1.383	1.885	
				df	1	2	
				Sig.	.240	.390	

结果基于每个最深处的子表中的非空行和列。

<sup>a</sup>. 卡方统计量在 .05 级别处有意义。

<sup>b</sup>. 该子表中超过 20% 单元格的期望单元格计数小于 5。卡方结果可能无效。

6.5 多响应变量的概念与分类

6.5.1 多响应变量的概念与分类

1. 多响应变量的概念

一般情况下，在实验研究中，每个被试对象在测定的尺度变量，尤其是名义变量或定序变量上有且只有一个测定值与之对应。但在调查研究中，尤其是问卷调查的多项选择题中，经常会遇到一名被调查者对一个问题的响应，不总是只有其中的一项，而是往往有多项，甚至是全部的情况。

例如，当问到“您喜欢什么颜色？”时，被调查者可能既喜欢红色，也喜欢蓝色和绿色。如果让被调查者按喜欢程度排顺序时，被调查者的回答是“红色第一，蓝色第二，绿色第三”。这就构成了对一个问题(变量)的多个选择(响应)。这种问题称作多项选择题(又称多选多项选择题)。这是在市场研究或许多领域对某事物评价的研究中经常会遇到的问题。这种在同一个问题中有多种结果可供多个选择的方式称为多重响应。描述多重响应的变量，就称为多重响应变量或多响应变量。

2. 多响应变量的分类与编码

多响应变量的分类取决于对问题的设计和对数据的整理及其数据文件的建立。

(1) 多响应二分变量集及其编码

在有关大众喜好服装颜色的调查中，下述问题的提问方式是典型的多项选择题：  
请您在下列喜欢的颜色编号上打“√”(可多选)。

①红色 ②橙色 ③黄色 ④绿色 ⑤青色 ⑥蓝色 ⑦紫色 ⑧黑色 ⑨白色

在这样的多项选择题中，由于应答者不只有一个选项上作出响应，为便于将他们的应答结果在 SPSS 中进行分析，一般在建立数据文件时，将这种题型的问题分解为多个二项选其一的单选题来处理，如上述问题可以等价地演变成表 6-7 所示的单选题形式来实现。

请您在选择服装时喜欢的主体颜色前的“□”中画“√”(可多选)。

表 6-7 服装颜色问卷

编 号	调 查 内 容	选 项	
1	您喜欢红色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
2	您喜欢橙色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
3	您喜欢黄色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
4	您喜欢绿色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
5	您喜欢青色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
6	您喜欢蓝色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
7	您喜欢紫色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
8	您喜欢黑色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否
9	您喜欢白色吗	<input type="checkbox"/> 是	<input type="checkbox"/> 否

当用变量来表示二项选一项的单选题的响应结果时，每个变量的值只能有表明“是”、“和”“否”的两个代码。这样的变量称为二分变量。在建立数据文件时，变量名使用相同的变量主名，后面加以不同序号组成，如本组问题的 9 个变量名是 color1~color9，以便分析和整理时识别。答案的编码规则为：回答“是”变量值为 1，回答“否”变量值为 0，其他值为缺失值。

由此，不难得出对多项选择题设置变量时，可将其分解成若干个最简单的单选题，要用与

题中选项数相等的变量数来存放该题的响应数据。显而易见，这些变量均为二分(值)变量，一般用值“0”表示未选该项，用值“1”表示选中该项。若要对该题调查结果进行完整描述，则需要将这些二分变量组合在一起形成一个新变量。故所谓的多响应二分变量集实际上是由若干个二分变量组成的变量集。

(2) 多响应分类变量集及其分析方法

在多项选择题中，研究者只对被调查者在选项上的分布情况感兴趣。在实际研究中，研究者更可能对被调查者对选项喜好的先后顺序感兴趣。

例如，作为服装主体颜色，您可以选择最喜欢的3种，在答案前的○中填写喜欢的顺序号(最喜欢的为①，其次为②、③)。

- 红    ○ 橙    ○ 黄    ○ 绿    ○ 青
- 蓝    ○ 紫    ○ 黑    ○ 白    ○ 说不清

在这个问题中，每个问题可以有3个答案。在建立数据文件时，要建立3个变量：color1～color3，表示答者按喜欢程度选择的3个颜色。答案变量的值均按填写的顺序值编码，即代码A表示选择红色、代码B表示选择橙色、C表示选择黄色、…、I表示选择白色、J表示说不清。例如，选择结果为①黑、②红、③蓝，则变量color1的值为H，变量color2的值为A，变量color3的值为F。当然也可以使用数字编码。

如果要被调查者对上述9种颜色(去掉“说不清”这个选项)整体按喜好程度进行排序，则需建立9个变量，每个变量有9个可能的回答，需要用9个代码表示。

这类问题与多项选择题类似，但有所区别，习惯上将其称为排序题，所要建立的变量数等于所要排序的项数。由于每个变量可取的编码数为选项数，通常都大于2，为与二分变量有所区别，称这样的变量为分类变量。因而，多响应分类变量集是由若干个分类变量组成的。每个分类变量都有两个以上的值作为回答者的答案。这些分类变量共同反映了被调查者对问题的看法，因此单个分析就会有失全面。

(3) 解决多响应问题的 SPSS 过程

无论是多项选择题还是排序题，由于每个大问题包含若干个子问题，在分析时如果使用单个变量进行分析肯定是不全面的，因此在 SPSS 中首先将每个题的若干答案组成一个综合变量即多重响应集，习惯上也称为多重响应集，然后对综合变量的各种取值进行分析。

对多重响应集的建立及其分析可在 SPSS 中通过【分析】的菜单项【多重响应】中的各项功能实现的，见图 6-12。

①【定义变量集】过程。用来定义并建立多响应二分变量集或多响应分类变量集。

②【频数】过程(注：汉化为“频率”，不妥)。对多响应二分变量集和多响应分类变量集进行频数分布分析。

③【交叉表】过程。对多响应二分变量集、多响应分类变量集与其他变量集或与原变量进行交叉表分析。

对多响应的二分变量集或多响应分类变量集也可以使用表格功能进行分析，即用【分析】菜单中的【(制)表】命令的【多重响应】过程(见图 6-1)来建立多响应二分变量集或多响应分类变量集，再利用表格功能中的统计分析功能得到更丰富的统计量和统计分析结果。

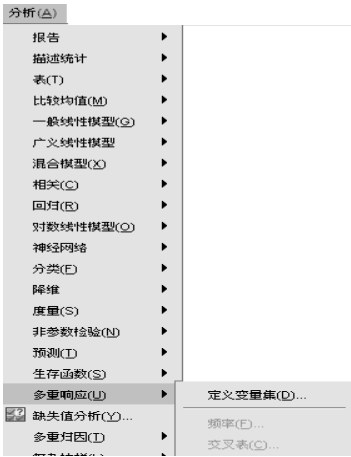


图 6-12 【多重响应】菜单项

值得一提的是，无论是【设定表】过程，还是【分析】菜单中的【多重响应】命令中的【频数】和【交叉表】过程均不支持对多重应答集进行显著性检验。

此外，在调用这两个过程中建立的多重响应集互不兼容，也即在表命令下建立的多重响应集只能在表命令下使用，在多重响应集命令下定义的多重响应集也只能在其下使用。尽管两者建立多重响应集的界面和操作过程几乎完全相似。为节省篇幅，在建立多重响应集中，下文只对在图 6-12 中所示的【定义变量集】过程作介绍。

6.5.2 定义与建立多响应变量集

定义多响应变量集是对多项选择题进行分析的必要步骤，必须把一组反映同一问题的多个答案变量组合在一个变量集中，方法如下：

(1) 按【分析→多重响应→定义变量集】顺序单击各菜单项，打开【定义多重响应集】对话框，如图 6-13 所示。

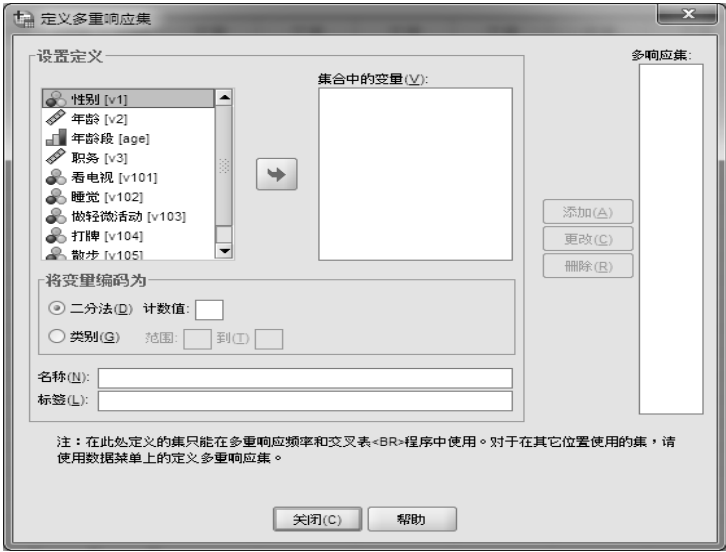


图 6-13 【定义多重响应集】对话框

(2) 在【设置定义】栏里选择同属于一个问题的多个答案变量，通过向右箭头按钮送入【集合中的变量】栏内，再根据该栏内的变量，定义变量集。

(3) 在【将变量编码为】栏内定义这组变量的编码方式。

①【二分法】。二分变量的计数值。如果所选择的变量是回答“是”、“否”的题目，选择此项，并在其后的编辑栏内输入想进行计数的答案代码；如要对回答“是”的观测进行计数，并且在数据文件中对每个问题选择“是”使用的代码为“1”，则在编辑栏内输入【1】。

②【类别】。分类变量。如果所选择的每个变量的回答是表示赞同顺序的数字，应该选择此项，并在其后的两个编辑栏内，输入要分析的变量的取值范围，即其值的起止范围。

(4)【名称】栏内为变量集命名。

(5)【标签】栏内输入变量集的标签。

(6) 单击【添加】按钮将定义好的变量名及其标签送入右面的【多响应集】栏内。该栏在命名的多重响应变量集前自动加“\$”以区别于一般变量。

(7) 反复进行上述操作，定义多个多重响应变量集。单击【关闭】按钮结束。

再次重申，使用以上功能菜单和上述方法定义的多重响应变量集，只能在图 6-12 所示的二级菜单的各项中使用，不可以用在【表】中构建自定义表格。按上述的方法定义的多重响应变量集可以对其进行频数分布分析和交叉表分析。

6.5.3 多响应变量的频数分布分析

多响应变量集的频数分布分析操作很简单，下面举例说明。

一项对公务员的营养、运动及心理调查中，除记录了被访者的性别、年龄外，还有如下问题和供选择的答案：

问题：您一般在晚饭后做什么？（可多选）

供选答案：A)看电视，B)睡觉，C)轻微活动，D)打牌，E)散步，F)其他(如看电影、跳迪斯科、加班、应酬等)。

6.5.3.1 多响应二分变量集的频数分布分析

**【例 2】** 根据答案建立 6 个变量 V101~V106，选择的变量值代码为“1”，未选择为“0”，见数据文件 data06-02。

按 6.5.2 节中介绍的方法建立多重响应变量集的标签为“晚饭后活动”。

1. 使用【多重响应】的【频数】进行频数分布分析的步骤

(1) 按【分析→多重响应→频数】顺序单击菜单项，打开【多响应频率】对话框，如图 6-14 所示。

(2) 在多响应频数分布分析对话框中，已经定义的变量集显示在左面【多响应集】栏中，选择要进行频数分布分析的变量集，本例只有一个，选中并送入右面的【表格】栏中。



图 6-14 【多响应频率】对话框

(3) 在【缺失值】栏中选择处理缺失值的方法。

- **【在二分集内按照列表顺序排除个案】**。将多响应二分变量集中任意一个变量值缺失的观测量从分析中剔除。只有当多响应变量集的所有组成变量是二分变量时才可以选择此项。系统默认的是只剔除多响应变量集中的所有变量都没有计数值的观测量。也就是说，观测量在多响应二分变量集中至少有一个变量包括计数值，该观测量就会计入频数分布表中。
- **【在类别内按照列表顺序排除个案】**。将多响应分类变量集中任意一个变量值不在定义范围内，被认为是缺失的观测量从分析中剔除。只有当多响应变量集的所有组成变量是分类变量时才可选择此项。默认的是当一个观测量的组成多响应分类集的所有变量没有一个变量的值包括在定义的范围时，该观测量才被认为是缺失的，要从分析中剔除。

显然，无论是二分变量集还是分类变量集，默认的处理方法，数据利用率较高。

2. 输出结果及说明(见表 6-8、表 6-9)

(1) 表 6-8 所示是观测量小结。有效观测量共 513 个，占 95.7%；缺失值观测量就是在二分变量集中的变量没有一个值是“1”的，都为“0”或者有“0”、“1”以外的值，这样的值有 23 个，占 4.3%。

- (2) 表 6-9 所示是多响应二分变量集中各变量的频数分布表。
- ① 组合的多响应变量集名为【\$rest】，标签为“晚饭后活动”。表中出现的都是变量标签。
- ② 标注【a】说明表中是计数值为“1”的频数。
- ③ 【响应】下的栏中是各变量响应的计数 N 和占回答为“1”的总数的百分比。
- ④ 表头为 N 的列对应左边变量值为 1 的发生频数，其总和 757，因为允许多选所以大于观测量总数 536(其中 23 个缺失值、513 个有效值)，757 为选择为“1”的总答案数。

表 6-8 观测量小结  
个案摘要

	个案					
	有效的		缺失		总计	
	N	百分比	N	百分比	N	百分比
a	513	95.7%	23	4.3%	536	100.0%

值为 1 时制表的二分组。

表 6-9 多响应二分变量集的频数分布表  
\$rest 频率

	响应		个案百分比
	N	百分比	
a 看电视	381	50.3%	74.3%
睡觉	57	7.5%	11.1%
做轻微活动	127	16.8%	24.8%
打牌	18	2.4%	3.5%
散步	174	23.0%	33.9%
总计	757	100.0%	147.6%

值为 1 时制表的二分组。

- ⑤ 【百分比】列说明 N 中的频数占总答案数 757 的百分比。总百分比为 100%。
- ⑥ 【个案百分比】列说明 N 中的频数占总观测量 513 的百分比。【总计】相当于 757 占总有效观测量数 513 的百分比，因此大于 100%。

从频数分布表中可以看到，晚饭后【看电视】的比例大大超过其他活动的比例，散步或做轻微活动的比例相对较少。这对健康是不利的，应该引起重视。

6.5.3.2 多响应分类变量集的频数分布分析

【例 3】 使用与例 1 相同的例题，如果问题是：按照您的习惯选择 3 个晚饭后的主要活动，并按经常性排列顺序。例如，最经常的是晚饭后看电视，其次是散步，有时打打牌，则应该在看电视前填写“1”，在散步前填写“2”，在打牌前填写“3”。

- ☐ 看电视
- ☐ 睡觉
- ☐ 轻微活动
- ☐ 打牌
- ☐ 散步
- ☐ 其他(继续工作、看电影、跳迪斯科或应酬等)

1. 建立数据文件  
见数据文件 data06-03。变量 vv1~vv3 分别表示第一选择到第三选择。
2. 对数据进行频数分布分析

- 频数分布分析的方式有两种：
- (1) 对 3 个变量分别进行频数分布分析
- ① 按【分析→描述统计→频数】(注：汉化为“频率”不妥)顺序打开【频数分布分析】对话框。
- ② 在主对话框中，将 vv1、vv2、vv3 这 3 个变量送入右面的【变量】栏，并选择【显示频率表格】选项(注：应为【显示频数分布表】)，要求输出频数分布表。
- ③ 单击【图表】按钮，在对话框的【图表类型】栏中选择【饼图】；在【图表值】栏选择表 6-10 观测量统计表

		第一选择	第二选择	第三选择
N	有效	532	497	274
	缺失	4	39	262

- 【百分比】，要求以百分比标注饼图的各分块。
- ④ 输出结果见表 6-10~表 6-13 和图 6-15~图 6-17。
- 表 6-10 所示是观测量小结，显示了每种选择的有效值和缺失值。

从表 6-11 第一选择的频数分布表和百分比饼 (见图 6-15) (注: 在饼图上加注百分比的做法: 在输出窗中双击饼图, 进入编辑该图的图表编辑窗, 单击右键, 在弹出的快捷菜单中选择【显示数据标签】, 即可将百分比值插在饼图上), 可以一目了然地看出:

- 52.4%的问卷回答者晚饭后活动的第一选择是“看电视”, 这是他们最经常的活动方式。
- 其次是散步 23.3%和轻微活动 16.8%。可见电视在人们的晚饭后活动中占有很重要的地位。“看电视”是大多数人首选的活动。“做轻微活动”的人和“散步”的人总和占 40.1%, 说明当前相当一部分人很重视晚饭后的活动。
- “打牌、睡觉和其他活动(或许是看电影、跳迪斯科、加班工作或应酬等)”的人数比例很少, 总和不到 10%。

表 6-11 第一选择频数分布表

第一选择		频数	百分比	有效百分比	累积百分比
有效	看电视	281	52.4	52.8	52.8
	睡觉	11	2.1	2.1	54.9
	轻微活动	90	16.8	16.9	71.8
	打牌	5	.9	.9	72.7
	散步	125	23.3	23.5	96.2
	其他	20	3.7	3.8	100.0
合计		532	99.3	100.0	
缺失	系统	4	.7		
	合计	536	100.0		

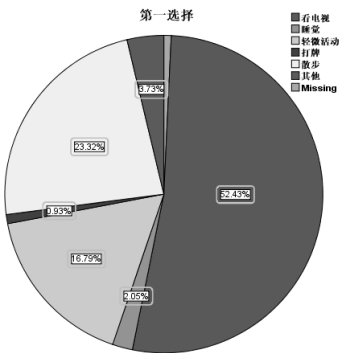


图 6-15 第一选择的饼图

表 6-12 第二选择频数分布表

第二选择		频数	百分比	有效百分比	累积百分比
有效	看电视	158	29.5	31.8	31.8
	睡觉	45	8.4	9.1	40.8
	轻微活动	70	13.1	14.1	54.9
	打牌	28	5.2	5.6	60.6
	散步	158	29.5	31.8	92.4
	其他	38	7.1	7.6	100.0
合计		497	92.7	100.0	
缺失	系统	39	7.3		
	合计	536	100.0		

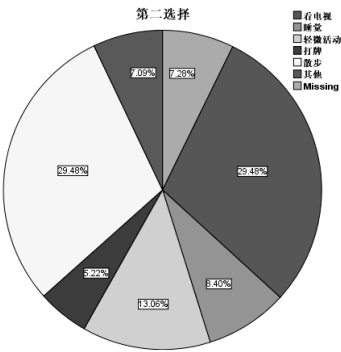


图 6-16 第二选择的饼图

表 6-13 第三选择频数分布表

第三选择		频数	百分比	有效百分比	累积百分比
有效	看电视	46	8.6	16.8	16.8
	睡觉	43	8.0	15.7	32.5
	轻微活动	68	12.7	24.8	57.3
	打牌	40	7.5	14.6	71.9
	散步	36	6.7	13.1	85.0
	其他	41	7.6	15.0	100.0
合计		274	51.1	100.0	
缺失	系统	262	48.9		
	合计	536	100.0		

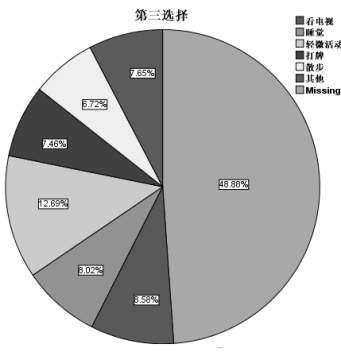


图 6-17 第三选择的饼图

从第二选择的频数分布表和百分比饼图可以看出，第二选择“看电视”和“散步”的百分比相当，“做轻微活动”的占有较大比例。

第三选择的频数分布表和百分比饼图表示将近一半的问卷回答者没有任何选择。也就是说，近 50%的人每天晚饭后经常安排两项活动，说明公务员业余生活比较单调。

如果想分析晚饭后看电视、散步等活动所占的总百分比就应该建立多响应分类变量集，并对变量集进行频数分布分析。

(2) 组成多响应分类变量集，对该变量集进行频数分布分析

① 定义多响应分类变量集。

- 按【分析→多重响应→定义变量集】顺序单击菜单项，打开对话框，将 vv1~vv3 送入【集合中的变量】框内。
- 在【将变量编码为】栏中选择【类别】选项，在其后输入 vv1~vv3 的取值范围，最小值“1”和最大值“6”。
- 在【名称】后面输入多响应分类变量集名“rest”，在下面一行输入标签“晚饭后活动”。单击【添加】按钮将变量名送入右面的【多响应集】栏内，单击【关闭】按钮。

② 进行频数分布分析。

- 按【分析→多重响应→频数】顺序单击菜单项，打开对话框，将多响应分类变量集【\$rest】送入【表格】栏。
- 单击【确定】按钮，在输出窗口中得到频数分布表；运行结果见表 6-14 和表 6-15。

表 6-14 观测量小结

	个案					
	有效的		缺失		总计	
	N	百分比	N	百分比	N	百分比
a	533	99.4%	3	0.6%	536	100.0%

组

表 6-15 多响应分类变量集的频数分布表

\$rest 频率			
		响应	
		N	百分比
a	看电视	485	37.2%
	睡觉	99	7.6%
	轻微活动	228	17.5%
	打牌	73	5.6%
	散步	319	24.5%
	其他	99	7.6%
总计		1303	100.0%

组

③ 结果解释。

表 6-14 是观测量小结，共 533 个有效观测量参与分析，占全部观测量的 99.4%；剔除缺失值 3 个，占全部观测量的 0.6%。

表 6-15 是多响应分类变量集的频数分布表。

- 左数第 1 列是组成多响应变量集的 3 个变量 vv1、vv2、vv3 共同使用的 6 个值标签。



- 第3列N是多响应分类变量取值1~6即各种晚饭后活动的总频数，也就是3个原始变量取各代码值的总计频数。
- 【总计】值，答案总数是1303。
- 第4列百分比中每个值是同行对应N值占选择答案总数1303的百分比。
- 第5列个案百分比中每个值是同行对应的N值占有效观测量总数的百分比。最后的百分比总和和自然会大于百分之百，因为每个回答者都有可能选择两个或3个答案。

根据第4列数据，即总回答数占选择答案总数的百分比可以看出，晚饭后经常看电视的人数占活动总数(答案总数)的37.2%，散步和轻微活动的百分比分别为24.5%和17.5%。说明这三项活动是公务员业余生活的主要内容。“轻微活动”与“散步”总数占41.9%比“看电视”的总百分比要大，说明公务员比较重视身体的活动。而其他活动(代码为6)只占7.6%，说明活动单调。

应该说明的是，该问题调查所使用的答案是经过初步调查设计的，答案很少，虽然排列了第一、第二、第三选择，由于大多数人的晚饭后活动确实单调，所以对3个变量的分析和对多响应分类变量集的频数分布分析的结果大体一致，没有体现出多响应变量频数分布分析的特点。这里仅作为一种方法加以介绍。

## 6.5.4 多响应变量的交叉表分析

### 6.5.4.1 多响应变量集交叉表分析过程

多响应变量集交叉表分析的步骤如下：

(1) 按【分析→多重响应→交叉表】顺序单击菜单项，打开【多响应变量集的交叉表】对话框，见图6-18，左上栏中显示了数据文件中所有数值型变量，下面栏中显示了定义好的多响应变量集。

(2) 可以选择多响应变量集作为行变量送入【行】变量栏，如果作为列变量则送入【列】变量栏，作为层变量则送入层变量栏。

(3) 可以选择基本变量作为交叉表的行、列、层变量送入相应的变量栏中。对基本变量，无论在交叉表中处于什么位置，应选择并送入相应的栏中后需要定义它们在交叉表中出现的取值范围。

(4) 在【行】、【列】、【层】栏中选择要定义分析范围的基本变量，单击【定义范围】按钮，打开【定义范围】对话框，如图6-19所示。在【最小值】栏中输入所选择变量的分类的最小值，在【最大值】栏中输入分类的最大值。最小值和最大值的选择可以不包括全部分类值。范围之内分类值将出现在交叉表中。



图 6-18 【多响应交叉表】对话框

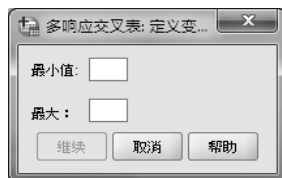


图 6-19 【定义范围】对话框

(5) 单击【选项】按钮打开【选项】对话框，如图 6-20 所示，指定输出选项。

① 【在单元格百分比】栏内选择交叉表的单元格内显示哪些统计量，可选择输出【行】百分比、【列】百分比、【总计】百分比。

② 【跨响应集匹配变量】。这是仅对多响应分类变量集可用的复选项。该复选项确定输出的交叉表中的对应关系。两个多响应分类变量集中的第一个集中的第一个变量与第二个集中的第一个变量作为一对，第一个集中的第二个变量与第二个集中的第二个变量作为一对，等等。如果选择此项，单元格中的百分比的基数是答案总数，而不是回答者的总数。因此该选项下面的【百分比基于】选项中只有【响应】是可以选择的，而且是默认的。

- ③ 在【百分比基于】栏中选择：
- 【个案(样品)】。交叉表中各单元格中的百分比的计算基数是观测量数，即答卷人数。
  - 【响应】。交叉表中各单元格中的百分比的计算基数是总响应数。由于是多项选择题，因此一般情况是总响应数大于应答者人数。

④ 【缺失值】栏选择项的含义与多响应变量集的频数分布中缺失值的含义相同，见第 6.5.3.1 节的相关内容。

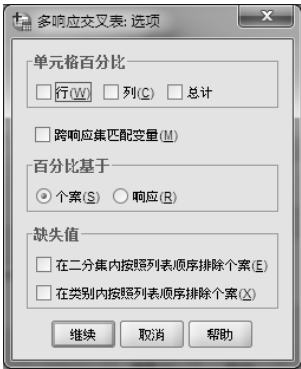


图 6-20 【选项】对话框

6.5.4.2 多响应二分变量集的交叉表分析实例

【例 4】 仍以数据文件 data06-02 的数据为例加以说明。分析 50 岁以下各年龄段的人饭后做什么。在数据文件中，原来只记录了年龄，为了分析，先根据年龄变量 V2 生成年龄段变量 age。利用【转换】菜单中的【计算变量】功能，按下列原则将年龄变量分段：

```
if V2<=17 then age=0; 不参与分析没有定义值标签，可以当作缺失值处理。  
if 17<V2<=30 then age=1; 值标签为“<=30”； 以下叙述中称之为 30 岁以下。  
if 30<V2<=40 then age=2; 值标签为“31~40”。  
if 40<V2<=50 then age=3; 值标签为“41~50”。  
if V2>50 then age=4; 值标签定义为“>50”。
```

1. 使用多响应变量分析功能进行交叉表分析

(1) 进行交叉表分析的步骤如下。

先定义多响应二分变量集\$rest，方法见 6.5.2 节，然后进行如下操作：

- ① 按【分析→多重响应→交叉表】顺序单击菜单项，打开【多响应交叉表】对话框。
- ② 做二维交叉表，选择年龄段变量 age 送入【行】变量栏。
- ③ 单击【定义范围】按钮，输入最小值为“1”，最大值为“3”。
- ④ 选择多响应二分变量【\$rest】送入【列】栏。
- ⑤ 单击【选项】按钮，在【单元格百分比】栏中选择【行】、【列】、【总计】选项，要求每个单元格除显示单元格频数外，都显示行百分比、列百分比和总百分比。

由于多响应变量集是由若干二分变量组成的，因此不能选择【跨响应集匹配变量】项。计算百分比的基数是观测数。选择【百分比基于】栏中的【个案】选项。

按【确定】按钮，运行后再次重复上述过程，但在交叉表【选项】对话框的【百分比基于】栏中选择【响应】选项。

(2) 输出结果见表 6-16～表 6-18。

(3) 结果解释。

表 6-16 所示为观测量小结，注意凡是没有包括在分析范围之内的观测量，例如小于 17 岁和大于 50 岁的都计数在缺失值内。因此有效值为 456，占总观测量的 85.1%，缺失值为 80 个。

表 6-17 所示是以观测量总数为百分比基数的交叉表。表中每个单元格中的数据自上至下为：该单元格的频数、行百分比、列百分比、总百分比。以左上角第一个单元格为例给出解释：该单元格表示 30 岁以下的人晚饭后看电视的有 63 人，占这个年龄段 93 人的 67.7%，占晚饭后看电视总人数(三个年龄段)337 人的 18.7%，占三个年龄段总人数 456 的 13.8%。

表 6-16 观测量小结表

个案摘要

	个案					
	有效的		缺失		总计	
	N	百分比	N	百分比	N	百分比
age*\$rest	456	85.1%	80	14.9%	536	100.0%

最右边一列的各单元格中，上边的数值是该行观测量总数，第二个数值是该行观测量总数占总观测量数的百分比。例如，年龄大于等于 18 岁、小于等于 30 岁的 93 人，注意因为允许多项选择，这个 93 并不等于该行中各单元格的计数之和。这个年龄段的 93 人占总人数 456 人的 20.4%。

表下面第一行【计数】为各列观测量总数，就是该列各单元格中计数之总和。【总计的%】是列观测量数占总观测量数的百分比。例如，晚饭后看电视的 337 人占观测量总数 456 的 73.9%。

右下角单元格中显示的是观测量总数 456，并表示以该数值为分母计算的各百分比。

注意：交叉表中的列变量是多响应二分变量集，因此最后右下角单元格中的总计数 456 不等于最后一行上各列答案计数之总和 670，而是总观测量数 456；最后一行上的各列百分比的总和自然也不是 100%，而是总答案数除以总观测量数的商，大于 100%，即  $670/456 = 146.9\%$ ，这个数在这里未列出。

比较表 6-17 和表 6-18 可以看出，各对应单元格中的频数是一样的，但是对应单元格中的行、列和总百分比是不同的。

表 6-18 是以答案总数为计算百分比基数的交叉表。以左上角单元格为例，小于 30 岁、大于等于 18 岁的人，晚饭后看电视的人数为 63 人，占这个年龄段回答者总人数(行总和)137 人的 46.0%(行百分比)，该频数占选择看电视总人数(列总和)337 人的 18.7%(列百分比)，占总答案数 670 的 9.4%。该频数占回答所在列表示的活动类型的人数的百分比和占总选择数的百分比。最右边一列中各单元格上边的数值是各行频数之和，下边的数值是各行频数之和占总答案数 670 的百分比。第一行为总频数，即总选择数 137，占 670 的 20.4%。

表最下面一行数值是各种选择(各列)的总频数，这与表 6-17 是相同的，而百分比却是各列总频数占总选择数 670 的百分比。

总答案数在右下角中是 670，它是各列总频数计数值之和。列百分比之和是 100%，同时也是行百分比之和。

表 6-17 以观测量总数为百分比基数的交叉表

age*\$rest 交叉制表								
		饭后活动 <sup>a</sup>					总计	
		看电视	睡觉	做轻微活动	打牌	散步		
年龄段	<=30	计数	63	9	27	5	33	93
		age 内的 %	67.7%	9.7%	29.0%	5.4%	35.5%	
		\$rest 内的 %	18.7%	18.4%	22.7%	29.4%	22.3%	
		总计的 %	13.8%	2.0%	5.9%	1.1%	7.2%	20.4%
	31~40	计数	127	19	42	5	51	170
		age 内的 %	74.7%	11.2%	24.7%	2.9%	30.0%	
		\$rest 内的 %	37.7%	38.8%	35.3%	29.4%	34.5%	
		总计的 %	27.9%	4.2%	9.2%	1.1%	11.2%	37.3%
	41~50	计数	147	21	50	7	64	193
		age 内的 %	76.2%	10.9%	25.9%	3.6%	33.2%	
		\$rest 内的 %	43.6%	42.9%	42.0%	41.2%	43.2%	
		总计的 %	32.2%	4.6%	11.0%	1.5%	14.0%	42.3%
总计	计数	337	49	119	17	148	456	
	总计的 %	73.9%	10.7%	26.1%	3.7%	32.5%	100.0%	

百分比和总计以响应者为基础。

a. 值为 1 时制表的二分组。

读者可以根据对各项的上述解释，观察交叉表，得出必要的结论。

表 6-18 以答案总数为百分比基数的交叉表

age*\$rest 交叉制表								
			a					总计
			看电视	睡觉	做轻微活动	打牌	散步	
年龄段	<=30	计数	63	9	27	5	33	137
		age 内的 %	46.0%	6.6%	19.7%	3.6%	24.1%	
		\$rest 内的 %	18.7%	18.4%	22.7%	29.4%	22.3%	
		总计的 %	9.4%	1.3%	4.0%	0.7%	4.9%	20.4%
	31~40	计数	127	19	42	5	51	244
		age 内的 %	52.0%	7.8%	17.2%	2.0%	20.9%	
		\$rest 内的 %	37.7%	38.8%	35.3%	29.4%	34.5%	
		总计的 %	19.0%	2.8%	6.3%	0.7%	7.6%	36.4%
	41~50	计数	147	21	50	7	64	289
		age 内的 %	50.9%	7.3%	17.3%	2.4%	22.1%	
		\$rest 内的 %	43.6%	42.9%	42.0%	41.2%	43.2%	
		总计的 %	21.9%	3.1%	7.5%	1.0%	9.6%	43.1%
总计	计数	337	49	119	17	148	670	
	总计的 %	50.3%	7.3%	17.8%	2.5%	22.1%	100.0%	

百分比和总计以响应为基础。

值为 1 时制表的二分组。

多响应分类变量集的交叉表分析的操作方法与多响应二分变量集的交叉表分析操作方法相同。读者可以自己实践，此处不再赘述。

注意：使用多响应变量集交叉表分析功能也可以作单个基本变量间的交叉表，但其功能不如【描述统计】二级菜单中的【交叉表】项的交叉表分析功能强。如果希望进行卡方检验，或得到表明分布情况的图形，分析原变量的频数分布和得到交叉表，还是应该使用【描述统计】二级菜单中的【交叉表】项的交叉表分析功能。但该处不支持多重响应集。

6.5.5 使用表功能分析多响应变量集

使用【分析】菜单的【表】功能也可以分析多响应变量集，要想得到很好形式的表格，可用【表】子菜单中的【设定表】功能做表并进行分析。

6.5.5.1 简单频数分布分析

- (1) 多响应二分变量集进行简单频数分析的操作步骤：
- 在数据文件 data06-02 的基础上，现已在【分析】菜单【表】命令的【多响应集】过程中建立了 V101~V106 的 6 个变量的一个名为“\$rest”、标签名为“晚饭后活动”的多重二分变量响应集。现按【分析→表→设定表】功能做表，对\$rest 进行频数分布分析。
- ① 打开【设定表】对话框，见图 6-4，把多响应变量集【\$rest】拖曳到【列】栏中。
  - ② 单击【标题】选项卡，在【标题】栏中输入表格标题“某单位公务员晚饭后活动类型频数分布表()”。光标置于括号中，单击【日期】按钮，要求显示分析日期。
  - ③ 在主对话框左下角【定义】栏中，单击【N%摘要统计量】按钮，打开相应对话框，见图 6-5，【统计量】栏中选择行【N%】送入右边的【显示】栏中。注意这是以总观测量数为基数计算百分比的。单击【应用选择】按钮，回到主对话框的【表】选项卡。
  - ④ 单击左下角的【分类和总计】按钮，打开相应对话框，见图 6-6。
- 将【显示】栏的各项前边的对钩去掉，要求不显示空单元，也不选择总计，因为对于多响应问题这里也显示总观测量数，而不是各项频数的总和，而各项频数的总和也没有什么实际意义故不选。单击【应用】按钮，返回主对话框。
- ⑤ 在主对话框中的【摘要统计量】栏中选择【行】选项，要求把统计量标题排在表格左边，每类统计量占一行。按【确定】按钮运行，结果见表 6-19。
- (2) 多响应分类变量集与二分变量集的操作步骤完全相同，见数据文件 data06-03。得到的表格是按前三种选择的频数分布，见表 6-20。

表 6-19 二分多响应集的频数分布分析结果

	晚饭后活动					
	看电视	睡觉	做轻微活动	打牌	散步	其他
计数	381	57	127	18	174	44
行 N %	71.8%	10.7%	23.9%	3.4%	32.8%	8.3%

表 6-20 多响应分类变量集的频数分布分析结果

	晚饭后活动					
	看电视	睡觉	轻微活动	打牌	散步	其他
计数	485	99	228	73	318	99
行 N %	91.0%	18.6%	42.8%	13.7%	59.7%	18.6%

6.5.5.2 交叉表分析

【例 5】仍以数据文件 data06-02 为例，使用【分析→表】中的【设定表】功能分析多响应变量的交叉表。

1. 多响应二分变量集的交叉表

- (1) 打开数据文件 data06-02，按【分析→表→多响应集】顺序打开定义多响应变量集的对话框，定义二分变量集\$rest。具体方法可参见 6.5.2 节。使用这个菜单定义的多响应变量集是可以保存的，到下一次打开 SPSS 时仍然存在，不用重新定义。

(2) 按【分析→表→设定表】顺序打开如图 6-4 所示的对话框, 将\$rest 拖曳到【列】栏中, 进行以下定义:

① 在主对话框左下角, 单击【N%摘要统计量】按钮, 在相应对话框的【统计量】栏中选择以下各常用项, 送入【显示】栏中:

- 【计数】。要求显示每个单元格的计数, 即每个年龄段, 晚饭后从事各种活动的人数。
- 【行 N %】。是以【计数】值作为分母的百分比。
- 【响应】。回答人数。数值上与【计数】值相等, 但总计、百分比有区别。
- 【行响应%】。各单元格中的回答人数相对于该行(同年龄段)总回答人数的百分比。
- 【表 N %】。该单元格的中的计数频数与样本量计数的百分比值。

单击【应用选择】按钮, 回到主对话框的【表】选项卡。

② 单击【分类和总计】按钮, 在相应的对话框(见图 6-6)的【显示】栏中选择【总计】要求显示总计, 去除【空类别】复选项, 要求不显示空单元。单击【应用】按钮, 将设置施加到\$rest 变量上。返回主对话框【表】选项卡。

(3) 将年龄段变量(age)拖曳到【行】栏中。

① 单击【分类和总计】按钮, 进入相应的对话框, 在【输出】栏中选择值为 3(标签 41~50)组, 在【小计和计算的类别】栏中单击【添加小计】按钮, 要求在该组后面加一个阶段总计。因为, 该研究更关心 50 岁以下公务员的饭后活动情况。在【显示】栏中只选择【总计】复选项, 要求显示总计。单击【应用】按钮, 将以上设置施加于【行】变量上, 返回主对话框【表】选项卡。

② 在主对话框的【摘要统计量】栏设置【位置】为行, 【源】设置为【列变量】, 即要求将列变量的综合统计量显示在行上。

(4) 运行该程序, 得到的结果见表 6-21。

可以看出, 使用表功能作为多响应变量的交叉表可以插入其他统计量, 如对 50 岁以下年龄的小计。它是对所在行以上的小结, 即 50 岁以下公务员的饭后活动的小结。选择“看电视”的人数为 337 人, 占 50 岁以下回答者总人数 472 人的 71.4%; 选择“轻微活动”和“散步”的人数为 267 人, 占 50 岁以下回答者总人数 472 人的 56.6%。相比之下, 选择“轻微活动”和“散步”的人比例不如看电视的大。

## 2. 多响应分类变量集的交叉表

**【例 6】** 以数据文件 data06-03 为例作多响应分类变量集的交叉表。

(1) 打开数据文件 data06-03, 按【分析→表→多响应集】顺序打开定义多响应变量集的对话框, 定义多响应变量集\$rest。具体方法可参见 6.5.2 节。

(2) 按【分析→表→设定表】顺序打开如图 6-4 所示的对话框, 将【\$rest】拖曳到【列】栏中。单击【分类和总计】按钮, 在相应的对话框(见图 6-6)中的【对类别排序】栏中选择【计数】, 【顺序】栏选择【降序】, 要求生成的表格按单元格降序排列; 在【显示】栏中只选择【总计】, 要求显示总计。单击【应用】按钮, 将设置施加到\$rest【行】变量上。返回主对话框【表】选项卡。

(3) 将 age 年龄段变量拖曳到【行】上。

① 单击【分类和总计】按钮, 进入相应的对话框, 在【输出】栏中选择 Value 值为“3”(标签 41~50)组, 在【小计和计算的类别】栏中单击【添加小计】按钮, 要求在该组后面加一个阶

段总计。因为，该研究更关心 50 岁以下的公务员饭后活动情况。在【显示】栏中只选择【总计】复选项，要求显示总计。单击【应用】按钮，将以上设置施加于【行】变量上，返回主对话框【表】选项卡。

表 6-21 用表功能实现交叉表

某单位公务员晚饭后活动类型频数分布表（2013/8/22）

			晚饭后活动						
			看电视	睡觉	做轻微活动	打牌	散步	其他	总计
年龄段	<=30	计数	63	9	27	5	33	10	97
		行 N %	64.9%	9.3%	27.8%	5.2%	34.0%	10.3%	100.0%
		响应	63	9	27	5	33	10	147
		行响应 %	42.9%	6.1%	18.4%	3.4%	22.4%	6.8%	100.0%
		表 N %	11.9%	1.7%	5.1%	0.9%	6.2%	1.9%	18.3%
	31~40	计数	127	19	42	5	51	21	180
		行 N %	70.6%	10.6%	23.3%	2.8%	28.3%	11.7%	100.0%
		响应	127	19	42	5	51	21	265
		行响应 %	47.9%	7.2%	15.8%	1.9%	19.2%	7.9%	100.0%
		表 N %	23.9%	3.6%	7.9%	0.9%	9.6%	4.0%	33.9%
	41~50	计数	147	21	50	7	64	11	195
		行 N %	75.4%	10.8%	25.6%	3.6%	32.8%	5.6%	100.0%
		响应	147	21	50	7	64	11	300
		行响应 %	49.0%	7.0%	16.7%	2.3%	21.3%	3.7%	100.0%
		表 N %	27.7%	4.0%	9.4%	1.3%	12.1%	2.1%	36.7%
	小计	计数	337	49	119	17	148	42	472
		行 N %	71.4%	10.4%	25.2%	3.6%	31.4%	8.9%	100.0%
		响应	337	49	119	17	148	42	712
		行响应 %	47.3%	6.9%	16.7%	2.4%	20.8%	5.9%	100.0%
		表 N %	63.5%	9.2%	22.4%	3.2%	27.9%	7.9%	88.9%
	>50	计数	44	8	8	1	26	2	59
		行 N %	74.6%	13.6%	13.6%	1.7%	44.1%	3.4%	100.0%
		响应	44	8	8	1	26	2	89
		行响应 %	49.4%	9.0%	9.0%	1.1%	29.2%	2.2%	100.0%
		表 N %	8.3%	1.5%	1.5%	0.2%	4.9%	0.4%	11.1%
	总计	计数	381	57	127	18	174	44	531
		行 N %	71.8%	10.7%	23.9%	3.4%	32.8%	8.3%	100.0%
		响应	381	57	127	18	174	44	801
		行响应 %	47.6%	7.1%	15.9%	2.2%	21.7%	5.5%	100.0%
		表 N %	71.8%	10.7%	23.9%	3.4%	32.8%	8.3%	100.0%

- ② 在主对话框中的【摘要统计量】栏设置【位置】为行，【源】设置为【列变量】，即要求将列变量的综合统计量显示在行上。
- (4) 将【性别】变量拖曳到【层】图标上，要求按性别分页做交叉表。单击【分类和总计】按钮进入相应的对话框，选择【显示】中的【总计】，其他不选择。单击【应用】按钮返回主对话框。
- (5) 运行程序，生成表 6-22 和表 6-23 所示的输出结果。

表 6-22 第一页交叉表

性别 男			晚饭后活动						
			其他	散步	打牌	轻微活动	睡觉	看电视	总计
年龄段	<=30	计数	12	19	12	24	16	43	51
	31~40	计数	21	47	12	33	15	82	89
	41~50	计数	15	76	15	46	8	97	102
	小计	计数	48	142	39	103	39	222	242
	>50	计数	6	33	4	17	4	37	40
	总计	计数	54	175	43	120	43	259	282

表 6-23 第二页交叉表

性别 女			晚饭后活动						
			其他	散步	打牌	轻微活动	睡觉	看电视	总计
年龄段	<=30	计数	1	21	5	18	20	42	46
	31~40	计数	23	49	7	36	20	79	91
	41~50	计数	18	59	16	45	12	87	94
	小计	计数	42	129	28	99	52	208	231
	>50	计数	3	14	2	9	4	18	20
	总计	计数	45	143	30	108	56	226	251

比较【表】菜单的【设定表】功能的表格和【多响应变量】分析中的表格，可以看出，【表】的制表功能更强，但是无论是【表】菜单的【设定表】还是【多响应变量】分析，对多响应变量集，无论是二分集还是分类集，都不能作任何检验，只能作各种百分比分析。

习 题 6

- 1. 用数据文件 data06-04 制表，表明不同性别、不同民族的平均工资，以职务等级作为层变量，自己定义表格标题。
- 2. 用数据文件 data06-04 将受教育程度重新分段编码：<=8 年的编码为 1；9 年~12 年的编码为 2；13~16 年的编码为 3；17 年以上的编码为 4。制表，表明不同受教育年限的各种职务的人数，不同受教育年限的各种职务的平均初始工资。性别作为层变量。
- 3. 多响应变量分几种类型？各对应哪种分析方法？
- 4. 在分析多响应变量之前，对原始数据应该进行怎样的处理？
- 5. 设计两种问卷对中央电视台的 10 个频道的认知情况进行调查，分别用两种方式建立数据文件，并对其知名度进行排序。

提示：

- 问卷 1：请说出你所知道的中央电视台的频道名称：顺序记录频道号和名称。
- 问卷 2：CCTV1 的频道名称你知道吗？答对记 1，不对记 0；CCTV2 的频道名称你知道吗？答对记 1，不对记 0。



# 第7章 基本统计分析

SPSS【分析】菜单下的【描述统计】中包括了一系列基本统计分析过程，如常用的频数分布表、描述统计量、交叉列联表及其独立性检验以及探索分析等。

当得到审核无误的原始数据后，需要认识数据、了解数据特征、检查数据分布，以便对数据作进一步判断，从而决定选择适合的统计分析方法分析数据。

针对离散型变量(也称分类变量)，可以使用频数分布分析，通过【频率】过程，对单一变量进行频数分布分析，认识变量的分布特征；通过【交叉表】过程可以对两个或两个以上的离散变量进行频数分布分析，实现二维及以上的交叉表、分层交叉表的分析，从而认识变量间的关联关系，并实现变量间的独立性检验。

针对连续型变量(即尺度测度变量)，可以使用【描述】过程计算描述统计量。反映数据集中性特征的统计指标称为集中趋势指标，如算术平均数、中位数和众数等；反映数据波动性特征的统计指标称为离中趋势指标，如全距(也称为极差)、平均差、方差和标准差等。另外，还可以计算偏度、峰度指标对变量的分布进行描述分析。

正态分布是连续型变量概率分布之一，在统计学中非常重要。它是中间分布频数多，两边分布频数少，均数为中心，左右对称的频数分布。变量服从正态分布是很多统计方法的前提条件。比如，正态或近似正态分布的变量可以使用均数、标准差进行简要描述分析，而非正态分布变量就不适合用平均数反映水平，而需要使用最大值、最小值、极差、中位数等进行分析。又比如，两个总体均数比较的T检验，其前提条件是变量服从正态分布，如果变量分布不呈正态，特别是小样本时，要使用非参数检验；在SPSS中可通过【探索】过程或【P-P图】和【Q-Q图】过程对变量是否服从正态分布进行初步验证。

SPSS的【描述统计】过程中还包括【比率】过程，用于对两个变量值比率变化的描述统计分析，适用于尺度变量资料。

## 7.1 频数分布分析

利用频数分布表可以对数据按变量值或按组进行分类整理，形成各变量的频数分布表和频数分布图，从而对各变量的分布特征有一个基本认识 and 了解，也可以通过该过程对数据进行审核和检查。

### 7.1.1 频数分布分析过程

(1) 建立或打开数据文件后，按【分析→描述统计→频率】顺序单击菜单项，打开如图7-1所示的对话框。

(2) 在源变量框中选择一个或多个变量，将其送入右侧的【变量】框中。

(3) 选中【显示频率表格】复选框，要求输出频数分布表。

(4) 单击【统计量】按钮，打开如图7-2所示对话框，选择要求输出的统计量。



图 7-1 【频率】主对话框

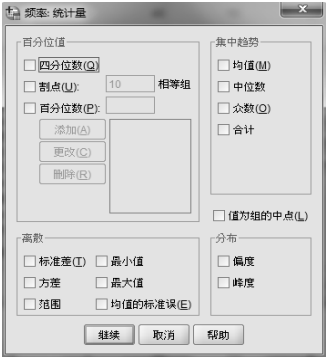


图 7-2 【频率: 统计量】对话框

- ① 【百分位值】栏。指定四分位数、百分位数等。
- 【四分位数】。输出四分位数，即第 25、50、75 百分位数。
  - 【割点】。输出等分点的百分位数。在参数框中可以输入 2~100 间的整数。例如，如果输入了“5”，即将数值按从小到大分为 5 等份，输出第 20、40、60、80 百分位数。
  - 【百分位数】。自定义输出百分位数。在参数框中输入 0~100 之间的数值，单击【添加】按钮可多次重复此操作，可指定输出多个百分位数。要剔除已定义的百分位数，只需选择它，单击【删除】按钮。
- ② 【集中趋势】栏。用于指定集中趋势指标。有以下复选项：【均值】、【中位数】、【众数】和【合计】。
- ③ 【离散】栏。用于指定离散趋势指标。
- 【标准差】。选中该项，输出标准差。
  - 【方差】。选中该项，输出方差。
  - 【最小值】、【最大值】和【范围】（作者注：软件翻译有误，此处指统计学中的全距，或称两极差）。输出的分别是最小值、最大值和全距。
  - 【均值的标准误】。输出均数的标准误。
- ④ 【分布】栏。指定描述数据分布的统计指标。
- 【偏度】。输出偏度值，并显示偏度的标准误。偏度值为“0”表明变量分布是对称的。
  - 【峰度】。输出峰度值及其标准误。峰度值为“0”说明变量分布是正态的。
- ⑤ 【值为组的中点】选项。是在计算百分位数和中位数时，假设数据已经分组，用各组的组中值代表各组数据。

(5) 在主对话框中单击【图表】按钮，打开如图 7-3 所示的【频率: 图表】对话框。在其中可设置统计图的类型及坐标轴等。

- ① 【图表类型】栏。选择统计图类型。
- 【无】。选中该项，不输出统计图，是系统默认状态。
  - 【条形图】。输出条形图，各条高度代表变量各分类的频数或百分比。不显示频数为 0 的分类。条形图适用于分类变量。
  - 【饼图】。输出饼图，不显示频数为 0 的类。适用于分类变量，整体构成为 1。
  - 【直方图】。仅适用于连续型变量。选择此项后，并选中【在直方图上显示正态曲线】，则显示的直方图中带有正态曲线。
- ② 【图表值】栏。在选择了条形图和饼图选项后生效。

- **【频率】**。直方图纵轴表示频数，饼图中的每个扇形表示该部分观测值的频数。
- **【百分比】**。直方图纵轴表示百分比，饼图的每个扇形为各观测值频数占总数的百分比。

(6) 在主对话框中单击**【格式】**按钮，打开如图 7-4 所示的**【频率：格式】**对话框。在其中可设置频数分布表输出格式。

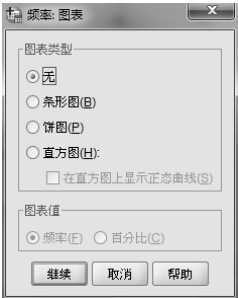


图 7-3 **【频率：图表】**对话框

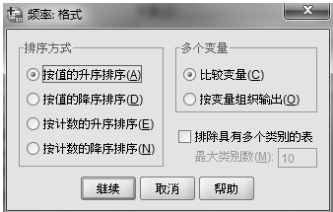


图 7-4 **【频率：格式】**对话框

① **【排序方式】**栏。设置频数分布表顺序，在选择了显示频率表格后生效。

- **【按值的升序排序】**。这是默认的排序方式。
- **【按值的降序排序】**。
- **【按计数的升序排序】**。
- **【按计数的降序排序】**。

如果设置了直方图或百分位数，频数分布表按变量值升序排列，而忽略用户的设置。

② **【多个变量】**栏。选择多变量输出表格。

- **【比较变量】**。所有变量的频数表集中输出。
- **【按变量组织输出】**。每一个变量单独输出一个频数表。

③ **【排除具有多个类别的表】**。控制频数表输出分类数。如果变量值的个数太多，占用空间，此时可以压缩它。默认值为 10，即如果变量值的个数大于 10，则不输出相应的频数分布表。

### 7.1.2 频数分布分析实例

**【例 1】** 数据文件 data07-01 为 1991 年美国社会调查数据。变量：race(种族)、happy(幸福感)。要求编制 race、happy 变量的频数分布表。

(1) 操作步骤

打开数据文件 data07-01，按**【分析→描述统计→频率】**顺序打开**【频率】**主对话框，在左侧框中分别选中**【种族 [race]】**、**【幸福感 [happy]】**变量，单击中间的箭头按钮，将其送入**【变量】**框中，确认选中**【显示频率表格】**选项，要求输出频数分布表，单击**【确定】**按钮，提交运行。

输出结果见表 7-1 和表 7-2。

(2) 结果解释

从表 7-1 中可看到，白种人 1264 人，占 83.3%；黑种人 204 人，占 13.4%；其他是 49 人，占 3.2%。没有缺失值。可见，本次调查中大部分是白种人。

从表 7-2 中可看到，有 13 个缺失值。被调查者中有 467 人感到非常幸福，有效百分比是 31.1%(分母是 1504)；872 人感到比较幸福，有效百分比是 58.0%；165 人感到不太幸福，有效

百分比是 11.0%。从结果看，有一半的被调查者感到比较幸福，有 89%的被调查者是幸福的(包括比较幸福和非常幸福)。

表 7-1 种族变量的频数分布表

种族		频数	百分比	有效百分比	累积百分比
有效	白人	1264	83.3	83.3	83.3
	黑人	204	13.4	13.4	96.8
	其他	49	3.2	3.2	100.0
	合计	1517	100.0	100.0	

表 7-2 幸福感变量的频数分布表

幸福感		频数	百分比	有效百分比	累积百分比
有效	非常幸福	467	30.8	31.1	31.1
	比较幸福	872	57.5	58.0	89.0
	不太幸福	165	10.9	11.0	100.0
	合计	1504	99.1	100.0	
缺失	NA	13	.9		
	合计	1517	100.0		

说明：本例中计算的两个变量，race 的测度类型是名义，happy 的测度类型是定序，这两种类型的数据适合使用频率过程进行频数分布分析。

【例 2】仍使用数据文件 data07-01。变量：age(年龄)、educ(受教育最高年限)。要求分析年龄和受教育最高年限的分布特征和描述统计量。

(1) 读取数据文件 data07-01，按【分析→描述统计→频率】顺序打开主对话框，选择【年龄[age]】和【受教育年数 [educ]】变量进入【变量】框中，选中【显示频率表格】选项，要求显示频数分布表。

(2) 单击【统计量】按钮，【百分位值】栏中选【四分位数】；【分布】栏中选【偏度】和【峰度】，检查数据的正态性；在【集中趋势】栏选【均值】和【中位数】；在【离散】栏选【标准差】、【最小值】、【最大值】和【范围(全距)】。单击【继续】按钮，返回主对话框。

(3) 单击【图表】按钮，打开【图表】对话框，选择【直方图和直方图上显示正态曲线】。单击【继续】按钮，返回主对话框。

(4) 在主对话框中，单击【确定】按钮，提交运行。

部分输出结果与分析见表 7-3、表 7-4 及图 7-5、图 7-6。

表 7-3 年龄与受教育年限变量的描述统计量

统计量		年龄	受教育最高年限
N	有效	1514	1510
	缺失	3	7
均值		45.63	12.88
中值		41.00	12.00
标准差		17.808	2.984
偏度		.524	-.168
偏度的标准误差		.063	.063
峰度		-.786	.710
峰度的标准误差		.126	.126
全距		71	20
极小值		18	0
极大值		89	20
百分位数	25	32.00	12.00
	50	41.00	12.00
	75	60.00	15.00

表 7-4 受教育年限变量的频数分布表

		受教育最高年限			
		频数	百分比	有效百分比	累积百分比
有效	0	2	.1	.1	.1
	3	5	.3	.3	.5
	4	5	.3	.3	.8
	5	6	.4	.4	1.2
	6	12	.8	.8	2.0
	7	25	1.6	1.7	3.6
	8	68	4.5	4.5	8.1
	9	56	3.7	3.7	11.9
	10	73	4.8	4.8	16.7
	11	85	5.6	5.6	22.3
	12	461	30.4	30.5	52.8
	13	130	8.6	8.6	61.5
	14	175	11.5	11.6	73.0
	15	73	4.8	4.8	77.9
	16	194	12.8	12.8	90.7
	17	43	2.8	2.8	93.6
缺失	18	45	3.0	3.0	96.6
	19	22	1.5	1.5	98.0
	20	30	2.0	2.0	100.0
	合计	1510	99.5	100.0	
	NA	7	.5		
	合计	1517	100.0		

表 7-3 为年龄和受教育年限变量的描述统计量。以年龄为例看输出结果，均值是 45.63，中位数是 41，二者相差较大，说明 **age** 变量是偏态的；偏度为 0.524，大于 0，说明 **age** 左偏，有一个较长的右尾，峰度值为-0.786，小于 0，曲线比较平缓。可以对照直方图认识这个变量。对于变量 **educ**，读者可以自己从输出表中认识它。

图 7-5、图 7-6 所示分别为 **age** 和 **educ** 变量带有正态曲线的直方图。这个正态曲线是按照数据的样本均值和标准差得到的。从图中可以比较明显地看到数据的分布与正态分布不一致，这与偏度、峰度值的结果一致。**age** 变量有一个较长的右尾，曲线较平缓；**educ** 变量右偏，左尾较长，曲线较陡峭。

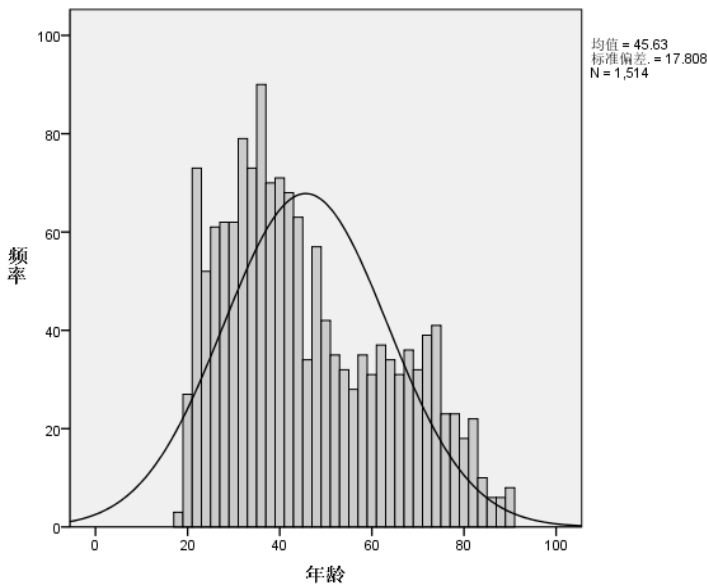


图 7-5 age 变量的直方图

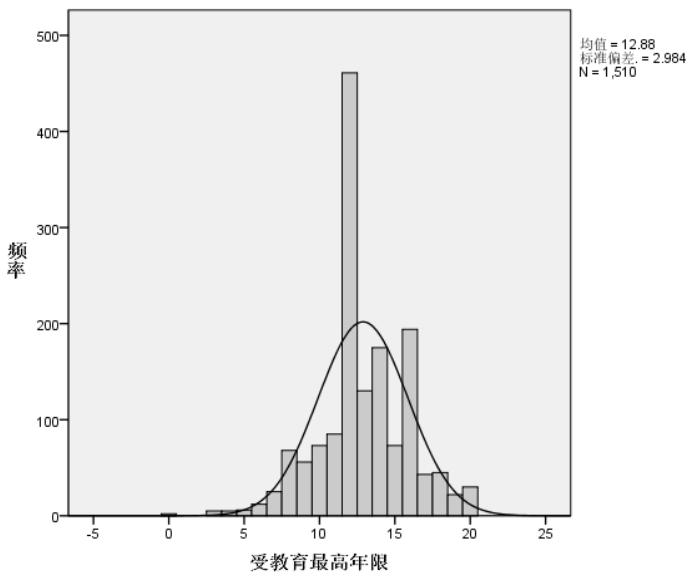


图 7-6 educ 变量的直方图

在这里需要说明的是:

- ① 图中正态曲线是根据变量的均值和标准差绘制的,不是标准正态分布。
- ② age 和 educ 变量的值都较多,最好先分组,然后再编制频数分布表。

## 7.2 描述统计

描述统计分析过程是通过计算均值、算术和、标准差、最大值、最小值、方差、全距和均数的标准误等统计量对变量进行描述,通过  $Z$  分数探明异常观测量。描述统计分析过程适用于尺度变量。至于使用哪些统计量作最终描述,要看其正态性检验的结果。

### 7.2.1 描述统计中的基本概念

#### 1. 均值、中位数和众数都是反映数据集中性特征的统计指标

当数据分布呈均匀分布或正态分布时,均值是反映一组数据平均水平常用的统计指标。当数据分布不对称或有极端值时,中位数是反映数据平均水平的一个较好的统计指标。在只有几个水平的分类变量时,可以使用众数,它是指一组数据中出现次数最多的那个数。如果中位数与众数相差很大,说明变量值中存在异常值。如果均值和中位数相差太大,说明数据的分布是偏态的。具体的计算公式参见第 8 章。

#### 2. 四分位数和百分位数都是描述变量值相对位置的统计指标

四分位数是指将一组数据按从小到大的顺序排序后,将其分成 4 等份,每个等分点上的值即为一个四分位数。百分位数是指将排序后的数据分成 100 等份,每个等分点上的值即为一个百分位数。较常用的百分位数是第 5 和第 95 百分位数。通过计算百分位数,可以了解某个值在集体中的位置。比如,收集到某个班 50 名学生的统计考试成绩,用该数据计算第 60 百分位数是 80 分,说明该班中有 60% 的学生成绩都在 80 分以下,有 40% 的学生成绩高于 80 分。四分位数实际就是百分位数中的第 25 百分位数、第 50 百分位数、第 75 百分位数,其中的第 50 百分位数也就是中位数。

#### 3. 极差、方差、标准差和标准误都是描述一组数据离散程度或变异大小的统计指标

具体的计算公式参见第 8 章。对正态分布数据常将均值和标准差结合在一起描述一组数据的特征。标准误是反映抽样误差大小的统计指标。标准误有均数的标准误和率的标准误,是样本均数(或样本率)的标准差。标准误是由样本均数(或样本率)推断总体均数(或总体率)可靠程度的统计指标。标准误小,说明样本均数(或样本率)与总体均数的差异小,抽样误差小,由样本均数(或样本率)推断总体均数(或总体率)的可靠性程度则高。

#### 4. 偏度和峰度是描述数据分布状况的统计指标

偏度,也称为偏斜度,描述数据分布的偏斜程度和方向。正态分布的偏度值为 0。偏度值为正值,分布左偏,右侧有长尾;偏度值为负值,则分布右偏,左侧有长尾。一个经验参考是,如果计算的偏度值在  $-1 \sim 1$  之间,则表明数据分布近似对称分布。峰度是描述数据分布曲线陡峭平缓程度的统计量。正态分布的峰度值是 0。如果峰度值为正,分布曲线比较陡峭,其峰比

标准正态分布的峰高，两端的尾部较长；如果峰度值为负，表明分布曲线是比较平缓的，其峰比标准正态分布的峰低，两端的尾部较短。

7.2.2 描述统计分析过程

描述统计分析过程主要计算数据的集中趋势和离中趋势指标。操作和内容如下：

(1) 按【分析→描述统计→描述】顺序打开【描述性】主对话框，如图 7-7 所示。

(2) 在源变量表中选择一个或多个变量作为待分析变量移入【变量】框中。

(3) 【将标准化得分另存为变量】。选中该项，则对所选择的每一个变量进行标准化，产生相应的 Z 分值，作为新变量保存在当前数据窗口中。其变量名为相应变量名加前缀 z。变量标准化的计算公式为

$$Z_i = \frac{x_i - \bar{x}}{s}$$

式中， $x_i$  为变量  $x$  的第  $i$  个观测值； $\bar{x}$  为变量  $x$  的均值； $s$  为变量  $x$  的标准差。

(4) 单击【选项】按钮，展开如图 7-8 所示的【描述：选项】对话框。在对话框中可以指定其他统计量与输出结果显示的顺序。

基本统计量参见 7.2.1 节中的介绍。



图 7-7 【描述性】主对话框

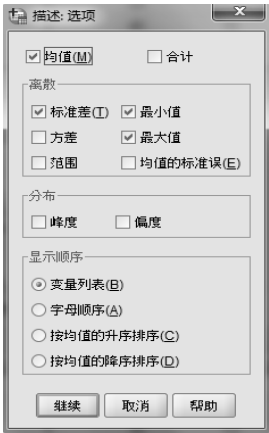


图 7-8 【描述：选项】对话框

7.2.3 描述统计分析实例

【例 3】 数据文件 data07-02 是对 1985 年美国联邦调查局对 50 个州各种犯罪情况调查的数据。变量 murder、rape、robbery、assault、burglary、larceny、autothft 分别为谋杀、强奸、抢劫、袭击、入室行窃、盗窃、盗车的案件数。对该数据进行描述统计分析。

(1) 打开数据文件，按【分析→描述统计→描述】顺序打开如图 7-7 所示的描述统计分析主对话框。

(2) 将 murder、rape、robbery、assault、burglary、larceny、autothft 变量送入【变量】栏中。

(3) 选中【将标准化得分另存为变量】，要求计算变量的标准化值，并保存到当前数据文件。

(4) 单击【选项】按钮，打开【选项】对话框。选中【均值】、【合计】、【标准差】、【最小值】、【最大值】、【范围(全距)】要求计算的描述统计量。

(5) 单击【继续】按钮返回主对话框。单击【确定】按钮提交运行，输出结果见表 7-5。

表 7-5 中，从左至右分别为变量名称、样本量、全距、最小值、最大值、算术和、均数及标准差。最后一行为有效样本量，本例是 50。

表 7-5 全美各种犯罪数据描述统计量

描述统计量							
	N	全距	极小值	极大值	和	均值	标准差
杀人事件	50	15	1	15	343	6.86	3.848
强奸事件	50	32	4	36	781	15.62	7.348
抢劫事件	50	437	7	443	5076	101.51	91.193
袭击事件	50	272	21	293	6771	135.42	68.170
入室行窃	50	1467	286	1753	46540	930.80	361.050
盗窃事件	50	2856	694	3550	97182	1943.64	709.829
盗车案	50	800	78	878	18393	367.86	199.610
有效的 N (列表状态)	50						

7.3 探索分析

7.3.1 探索分析的意义和数据要求

1. 探索过程提供对测得数据在分组与不分组的情况下的审核与观察

审核与观察可以有以下两个方面：

(1) 检查数据是否有错误

过大或过小的数据均有可能是异常值、影响点或是错误输入数据。对于这样的数据第一要找出，第二要分析原因，第三要决定是否从后续的分析中剔除。因为异常值和影响点往往对分析结果影响较大，不能真实地反映数据的总体特征。

(2) 验证变量分布特征

许多分析方法对数据的分布有要求。检查数据是否为正态分布以便选择分析方法。

另外，对若干组数据均值差异性的分析需要根据其方差是否相等，选择进行检验的计算公式。所以，分析之前需要首先要验证其方差的齐次性等。

2. 探索过程对变量和数据的要求

探索过程要求参与分析的变量是等间隔测度的数值型变量。分类变量是数值型或是字符型。箱图中用来标识异常值的变量可以是字符型也可以是数值型。

3. 探索过程提供观察和验证变量的方法

探索过程除了输出描述统计量外，还提供图形可以直观地将异常值、极端值、缺失数据及数据本身的特点表现出来。同时提供正态性检验，为选择分析方法提供依据。

(1) 箱图(见图 7-9)

它是对任何分布的数据的整体描述。

① 矩形框是箱图的主体，上、中、下三条线分别表示变量值的第 75、50、25 百分位数。

② 中间的纵向直线称为触须线。上截止横线是变量的最大值，下截止横线是变量的最小值。



③ 定义四分位数间距(IQ)，它是第 75 百分位数与第 25 百分位数之差。异常值使用“O”标记，分两种。箱体上方的 O 标记的变量值是大于  $U_1$ (第 75 百分位数+1.5IQ) 的值；箱体下方 O 标记的变量值是小于  $L_1$ (第 25 百分位数-1.5IQ) 的值。

④ 极端值使用“\*”标记。上极端值是大于  $U_2$ (第 75 百分位数+3IQ) 的值。下极端值是小于  $L_2$ (第 25 百分位数-3IQ) 的值。

(2) 茎叶图(见图 7-10)

茎叶图直观地描述数据的频数分布。可以自左至右分为三大部分：频数、茎、叶。茎表示数值的整数部分，叶表示数值的小数部分。每行的茎和每个叶组成的数字相加再乘以茎宽为茎叶所表示实际数据的近似值，即近似值=(茎值+叶值×0.1)×茎宽。

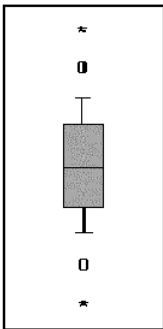


图 7-9 箱图

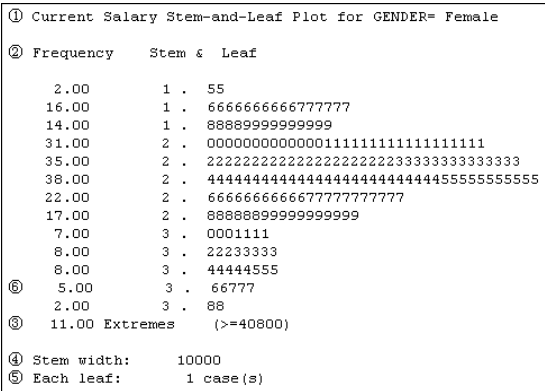


图 7-10 茎叶图

图 7-10 所示是 Current Salary 变量中 gender = female 的观测量的茎叶图。自左向右分别为频数、茎 Stem、叶 Leaf。图 7-10 中标有“③”的一行表示存在 11 个极端值，大于等于 40800；标有“④”的一行说明茎的宽度为 10000；标有“⑤”的一行中每一个叶代表 1 个观测量。

以“⑥”这一行为例，茎为 3，频数为 5。这行叶的组成为 6、6、7、7、7。按照观测量的近似值=(茎+叶×0.1)×茎宽的公式，第 1 个观测量的值为(3+0.6)×10000=36000。依此类推，这一行 5 个 Salary 变量的值近似为 36000、36000、37000、37000、37000。

(3) 正态性检验

除偏度、峰度观察变量分布特征外，探索过程还提供正态性验证方法及 P-P 图或 Q-Q 图观察。正态性检验同时输出 Kolmogorov-Smirnov 统计量(简称 K-S 统计量)和 Shapiro-Wilk 统计量。前一种检验建立在样本分布与预期累积分布之间没有显著差异假设的基础上，是用 Lillifors 显著性概率进行修正检验正态性的。Lillifors 可以在方差与均值未知的情况下直接使用，它是对 K-S 统计量的修正。如果指定的是非整数加权，加权样本量在 3~50 之间时，计算 Shapiro-Wilk 统计量；对于未加权或整数加权，当加权样本量在 3~5000 之间时计算 K-S 统计量。显著水平 Sig<0.05 时，拒绝正态分布假设。

(4) 方差齐性检验

许多检验要求方差齐性。例如，方差分析要求各分组样本的数据来自方差相同的正态总体；在进行独立样本 T 检验之前也需要事先确定两组方差是否相同；在进行多个均值组间比较时，也需要对方差是否相等进行选择。

如果各组方差不等，可以对数据进行转换来稳定方差或使方差尽可能地相等。

① 展布对水平图。(Spread vs. Level)用来判断各组离散程度是否相同。显示图形的同时,还输出回归方程斜率以及对数据进行幂转换的幂值。它们之间的关系为:幂值=1-回归斜率。如果没有指定因素变量,则不生成此图。

② Levene 检验。该方法最大的好处是对两个样本的数据进行方差齐性检验时,不强求数据必须服从正态分布。Levene 检验法先计算出离均差(各观测值减组均值的差),然后再通过离均差的绝对值进行单因素方差分析。如果显著水平值小于 0.05,则拒绝各组方差相等的假设。如果选择了数据转换,Levene 检验是根据转换后的数据计算的。

在进行方差齐性检验时,SPSS 提供了 4 种指标进行判断,分别是依据均值、依据中位数、依据中位数与调整后的自由度、依据调整的均值所得的各个统计量。

这几种统计量各有利弊,均值容易受到最大值、最小值以及极端值的影响;后三种都是比较好的方法,但它们都是将极端值排除在外,调整均值则排除了一部分观测量数据。

③ M-估计量。即集中趋势最大似然比稳健估计统计量。它是样本数据均值与中位数统计量的另外一种表现形式。当数据的分布具有较长尾部或者具有极值时,M 估计统计量要比均值以及中位数给出更精确的结果。

M 估计统计量在计算时对所有观测量加权。权重随观测量距离分布中心的远近而变,计算时包括极端值。极端值由于靠外,因此比位于中心部位的观测量给予的权重较小。M 估计不要求变量值呈正态分布。当数据分布均匀并且两尾较长或者当数据中存在极端值时,M 估计可以给出比均值或者中位数更合理的估计。

M 估计有 Huber、Andrew、Hampel 和 Tukey 估计方法。实践说明,这 4 种方法都可以很好地取代平均值以及中位数,其中 Huber 估计方法对于近似正态分布的数据效果最好。

### 7.3.2 探索分析过程

(1) 按【分析→描述统计→探索】顺序,打开如图 7-11 所示的【探索】主对话框。

(2) 从源变量框中选择若干个数值型变量作为因变量送入【因变量列表】框中。此时单击【确定】按钮即可获得默认的统计分析,这其中包括箱图、茎叶图以及描述统计量。默认情况下缺失值将会被排除到分析过程之外。

(3) 指定分组变量。在源变量框中选择一个或多个分组变量进入【因子列表】框中。分组变量可以将数据按该变量中的观测值进行分组分析。如果选择的分组变量不止一个,那么会以分组变量各取值进行组合分组。例如,指定分组变量:性别 sex(f、m)、年龄段 age(11、12、13),则按组合分组(f,11)、(f,12)、(f,13)、(m,11)、(m,12)、(m,13)对数据进行分析。

(4) 选择标识变量。在源变量表中指定一个变量作为观测量的标识变量,送入【标注个案】框中。

(5) 【输出】栏。确定输出项。

① 【两者都】。输出图形以及描述统计量。

② 【统计量】。只输出描述统计量。

③ 【图】。只输出图形。



图 7-11 【探索】分析过程主对话框

(6) 选择描述统计量。单击【统计量】按钮，打开如图 7-12 所示的对话框。

①【描述性】。系统默认选项，输出的描述统计量有：均值、中位数、众数、5%的调整平均值、标准误、方差、标准差、最大值、最小值、全距、四分位数、偏度及其标准误、峰度及其标准误。

【均值的置信区间】选项，计算均值的置信区间。在参数框中输入置信水平，选择的范围为 1%~99%，常用的数值为 90%、95%、99%，95%为默认值。

②【M-估计量】。输出集中趋势最大似然比的稳健估计。

③【界外值】。输出 5 个最大值与最小值，在输出窗口中它们被标明为极值。

④【百分位数】。输出第 5、10、25、50、75、90 以及 95 百分位数。

(7) 统计图形及其参数的选择。打开【探索：图】对话框，见图 7-13。

①【箱图】栏。通过该栏内容可以绘制箱图。

- 【无】不输出箱图。

- 【按因子水平分组】。因变量按因子变量分组，各组箱图并列输出。

- 【不分组】。所有因变量在一个图形中生成各组箱图，利于比较。

②【描述性】栏。选择描述统计分析图形。【茎叶图】是默认选项，生成茎叶图；【直方图】，生成频数分布统计图。

③【带检验的正态图】。输出 K-S 统计量及其 Lilliefors 显著性概率、Shapiro-Wilk 统计量及其显著性概率和 Q-Q 图。

④【伸展与级别 Levene 检验】(应为【带 Levene 检验的展布对水平图】)栏。输出展布对水平图(在 SPSS 输出中，将该图标注为“分布与水平图”)，同时输出回归直线斜率以及方差齐性的 Levene's 检验结果。如果没有指定分组变量，此选项无效；如果选择了【已转换】选项，将依据转换后的数据进行计算。

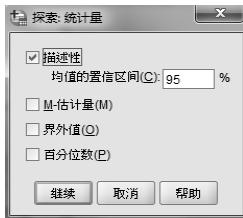


图 7-12 【探索：统计量】对话框

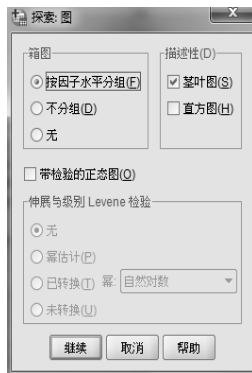


图 7-13 【探索：图】对话框

- 【无】。不产生展布对水平图，不进行方差齐性的 Levene 检验，是默认选项。

- 【幂估计，估计幂值】。对每一组数据产生一个中位数的自然对数与四分位数的自然对数的展布对水平图。同时为了使每组中的数据方差相等对数据进行幂变换。这个结果常常用来确定转换时最合适的幂值。

- 【已转换】。对原始数据进行转换，由读者在【幂：】框中指定幂转换使用的幂值。可以指定的幂值有：自然对数、1/平方根、倒数、平方根、平方、立方。

- **【未转换】**。不对数据进行转换。
- (8) 单击**【选项】**按钮，展开如图 7-14 所示的对话框。选择分析过程中对缺失值的处理方式。
- ① **【按列表排除个案】**。剔除带有缺失值的观测量，这是默认选项。
- ② **【按对排除个案】**。成对剔除有缺失值的观测量。
- ③ **【报告值】**。分组变量中的缺失值将被单独分为一组，显示在频数表中。

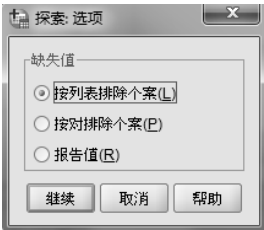


图 7-14 **【探索：选项】**对话框

7.3.3 探索分析实例

**【例 4】** 数据文件 data07-03 是 1969—1971 年美国一家银行的 474 名雇员情况的数据，包括变量：salary 当前薪水、educ 受教育年限(年)、prevexp 工作经历(月)、minority 是否是少数民族(0：非少数民族，1：少数民族)、jobcat 工作分类、id 雇员序号等。下面以 salary 当前薪水变量为例，说明探索分析的操作过程及其结果。

- (1) 选择变量，指定选项。
    - ① 打开数据文件 data07-03，按**【分析→描述统计→探索顺序】**，打开探索分析主对话框。
    - ② 选择**【薪水[salary]】**变量进入**【因变量】**列表框，选择**【性别[gender]】**变量进入**【因子】**列表框，选择雇员序号[id]变量进入**【标注个案】**框。在**【输出】**栏中，选择**【两者都】**项。
    - ③ 单击**【统计量】**按钮，打开对话框。选中**【描述性】**、**【M-估计量】**、**【界外值】**。
    - ④ 单击**【绘制】**按钮，打开**【探索：图】**对话框。**【箱图】**栏中选择**【按因子水平分组】**，**【描述性】**栏中选中**【茎叶图】**，选中**【带检验的正态图】**，在**【伸展与级别 Levene 检验】**栏中选择**【幂估计】**。单击**【继续】**按钮，返回主对话框。
    - ⑤ 在主对话框中单击**【确定】**按钮，提交运行。
  - (2) 部分输出结果见表 7-6～表 7-11 及图 7-16～图 7-19，重点进行解释。
- 表 7-6 所示是分析变量的描述统计量(注：作者将一个表分成两个表显示。)：因变量薪水、分组变量性别、四分位距是四分位数之差，其他统计量说明略。女雇员的薪水偏度值为 1.863，峰度值为 4.641，说明变量薪水的分布不呈正态。

表 7-6 Salary 的描述统计量

描述				描述					
性别			统计量	标准误	性别			统计量	标准误
薪水	女	均值	\$26,031.92	\$514.258	薪水	男	均值	\$41,441.78	\$1,213.968
		均值的 95% 置信区间	下限	\$25,018.29			均值的 95% 置信区间	下限	\$39,051.19
				上限					\$27,045.55
		5% 修整均值	\$25,248.30				5% 修整均值	\$39,445.87	
		中值	\$24,300.00				中值	\$32,850.00	
		方差	57123688.27				方差	380219336.3	
		标准差	\$7,558.021				标准差	\$19,499.214	
		极小值	\$15,750				极小值	\$19,650	
		极大值	\$58,125				极大值	\$135,000	
		范围	\$42,375				范围	\$115,350	
		四分位距	\$7,012				四分位距	\$22,675	
		偏度	1.863	.166			偏度	1.639	.152
		峰度	4.641	.330			峰度	2.780	.302

表 7-7 中的 a、b、c、d 分别表示 4 种 M 估计统计量，它是根据各自的加权常数计算的。与表 7-6 的均值比较，发现 M 估计值全部比均值小(Female=\$26031.92，Male=\$41,441.78)，且相差较大，据此可初步判定各组数据不是来自正态分布总体。M 估计值与中位数(女：\$ 24300；男：\$ 32850)十分接近。

表 7-8 中的“案例号”是观测样品的编号，“雇员序号”是 id。显示了按性别分组的各组中的 5 个最大值(最高薪水)和 5 个最小值(最低薪水)。

表 7-7 M 估计量

M-估计量				
性别	Huber 的 M-估计器 <sup>a</sup>	Tukey 的观权重 <sup>b</sup>	Hampel 的 M-估计器 <sup>c</sup>	Andrews 波 <sup>d</sup>
薪水 女	\$24,606.10	\$24,015.98	\$24,419.25	\$24,005.82
薪水 男	\$34,820.15	\$31,779.76	\$34,020.57	\$31,732.27

a. 加权常量为 1.339。  
b. 加权常量为 4.685。  
c. 加权常量为 1.700、3.400 和 8.500  
d. 加权常量为 1.340\*pi。

表 7-9 所示为检验数据是否为正态分布的统计量。自左至右分别为：Kolmogorov-Smirnov 统计量值、自由度、显著性概率，Shapiro-Wilk 检验的统计量、自由度、显著性概率。因为表中 Kolmogorov-Smirnov 下的显著性概率值为 Sig=0.000<0.05，所以拒绝数据呈正态分布的假设。

表 7-10 所示为方差齐性检验结果。自左至右分别为：Levene 统计量、自由度 1、自由度 2 和显著概率值；自上至下分别为：基于均值、基于中值、基于中值和带有调整后的自由度、基于修整均值所得的各个统计量。依据各种集中趋势统计量所作检验的显著概率值全部低于 0.001，拒绝方差相等的零假设，即男、女薪水方差不具有齐次性。

表 7-8 变量的极端值

极值			案例号	雇员序号	值
薪水 女	最高	1	371	371	\$58,125
		2	348	348	\$56,750
		3	468	468	\$55,750
		4	240	240	\$54,375
		5	72	72	\$54,000
	最低	1	378	378	\$15,750
		2	338	338	\$15,900
		3	411	411	\$16,200
		4	224	224	\$16,200
		5	90	90	\$16,200
男	最高	1	29	29	\$135,000
		2	32	32	\$110,625
		3	18	18	\$103,750
		4	343	343	\$103,500
		5	446	446	\$100,000
	最低	1	192	192	\$19,650
		2	372	372	\$21,300
		3	258	258	\$21,300
		4	22	22	\$21,750
		5	65	65	\$21,900

表 7-9 正态分布检验结果

正态性检验						
性别	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	统计量	df	Sig.	统计量	df	Sig.
薪水 女	.146	216	.000	.842	216	.000
男	.208	258	.000	.813	258	.000

a. Lilliefors 显著水平修正

表 7-10 方差齐性检验结果

方差齐性检验					
		Levene 统计量	df1	df2	Sig.
薪水	基于均值	119.669	1	472	.000
	基于中值	51.603	1	472	.000
	基于中值和带有调整后的 df	51.603	1	310.594	.000
	基于修整均值	95.446	1	472	.000

图 7-15(a)、(b)所示分别为男、女工资水平的茎叶图，从图中可以推断男雇员工资集中在 25000~39000 之间，女雇员工资集中在 16000~29000 之间，可以看出男女之间的工资水平可能有较大差异。

[illegible]

图 7-15 按性别变量分组的薪水茎叶图

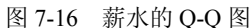


图 7-17 所示为性别变量 **gender** 的两个分组的薪水箱图。女雇员当前工资水平的全距较男雇员小。两组变量中都存在不少异常值, 如男雇员中的 29、32、18、103 号观测值, 女雇员中的 413 号观测值等。男雇员中的 29 号、女雇员中的 80、348 号观测值都是极端值。查看这些观测和其他变量值, 分析原因, 可确定后续的分析中是否包括这些观测或按其他变量分组分析。

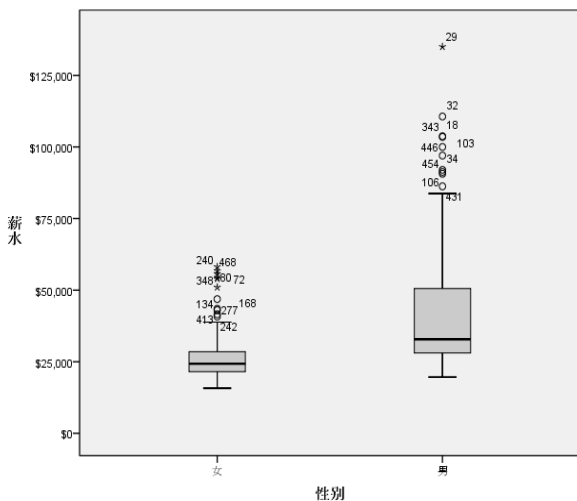


图 7-17 薪水的箱图

## 7.4 交叉表分析

SPSS 中的交叉表过程可以生成二维或多维(分层)分类变量行列交叉的频数表,可以计算分类变量之间的关联程度,并可以进行分类变量之间关联关系的独立性检验。

### 7.4.1 交叉表及其独立性卡方检验的思路

## 1. 交叉表的概念

在实际分析问题中，常常将两个分类变量联系起来讨论它们之间是否存在关联，如收入高低和地区之间是否存在关联，收入高低与性别之间是否存在关联，性别与是否喜欢体育锻炼之间是否存在关联等。对于这类问题的研究在统计学中可以使用列联表将两个问题联系起来进行描述。一个变量作为行变量，其值的个数  $r$  即为行数；另一个变量作为列变量，其值的个数  $c$  即为列数，形成  $r \times c$  交叉表(也称列联表)。最简单的交叉表是  $2 \times 2$  的四格表。例如，性别(男、女)与是否喜欢体育锻炼(喜欢、不喜欢)两个变量的关联性分析就可以通过一个  $2 \times 2$  的四格表形式进行描述。又如，表 7-11 所示的百货店与服务满意度交叉列联表就是一个  $4 \times 5$  的列联表。表中的数值是符合行列交叉情况发生的频数。

表 7-11 百货店与服务满意度交叉表

百货店\*服务满意度交叉列联表

		服务满意度					合计
		非常不满意	有些不满意	一般	比较满意	非常满意	
百货店	第一百货店	25	20	38	30	33	146
	第二百货店	26	30	34	27	19	136
	第三百货店	15	20	41	33	29	138
	第四百货店	27	35	44	22	34	162
合计		93	105	157	112	115	582

## 2. 交叉表独立性检验基本思路

在统计学中可以通过交叉表的独立性检验对两个变量是否存在关联进行分析。该方法的基

本思想与假设检验的基本思想是一样的。首先建立一个无效假设，即认为两个事物之间是独立的，没有关联。在假设成立的前提下，建立一个 $\chi^2$ 统计量，并计算它发生的概率，根据小概率事件在一次试验中不可能发生的原理，判断建立的无效假设是否成立。若拒绝无效假设，则做出两个事物之间存在关联的判断。因此交叉表的独立性检验也称为交叉表的 $\chi^2$ 检验。 $\chi^2$ 统计量的公式为

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

式中， $A$ 是实际频数； $T$ 是期望频数。

**注意：**使用这个统计量公式进行检验时，要求期望频数大于等于 5。若不满足该条件需要使用精确检验法。

7.4.2 交叉表分析过程

一个行变量和一个列变量可以形成一个二维交叉表，再指定一个分组变量作为控制变量（也称层变量）就形成三维交叉表。如果可以指定多个行、列、控制变量，就会形成复杂的多维交叉表。交叉表的变量可以是数值型、字符型或短字符串变量。

- (1) 按【分析→描述统计→交叉表】顺序打开如图 7-18 所示的【交叉表】主对话框。
- (2) 在源变量框中选择一个或多个分类变量送入【行】框，作为交叉表中的行变量。
- (3) 在源变量框中选择一个或多个分类变量送入【列】框，作为列变量。
- (4) 选择一个层变量进入【层 1 的 1】框中。单击【下一张】按钮，可指定另外一个控制变量；单击【上一张】按钮可改变前一次确定的控制变量。
- (5) 【显示复式条形图】。显示各组中各变量的分类条形图。
- (6) 【取消表格】。只输出统计量，不输出交叉表。
- (7) 单击【统计量】按钮，打开【交叉表：统计量】对话框，如图 7-19 所示。



图 7-18 【交叉表】主对话框



图 7-19 【交叉表：统计量】对话框

- ① 【卡方】。输出 3 种卡方检验结果。
- 【皮尔逊卡方检验】。检验的假设是行、列变量相互独立。当自由度大于 1、单元格频数大于 5 时，检验效果较好，是常用的检验方法。



- **【似然比卡方检验】**。对数线性模型检验方法之一，也是拟和优度检验方法。当样本量较大时，该统计量服从卡方分布。
- **【Fisher 的精确检验】**。当样本数小于 20 或单元格期望值有小于 5 时，使用该方法。
- **【线性与线性组合】**。检验假设是行、列变量相互独立，适合两个变量是定序或尺度变量的数据。

② **【相关性】**。输出 Pearson 和 Spearman 相关系数。分别表示两变量的线性相关或变量秩之间的关联程度。数值范围为-1~+1，0 表示无线性关系，符号表示相关方向。如果行、列变量为定序变量，应计算 Spearman 相关系数。如果两个变量是分类变量，不适合使用该选项计算关联程度。

③ **【Kappa】**。输出 Cohen 的 Kappa 系数。用来检验对同一对象两种评估的一致性，它仅适用于具有相同分类值和相同分类数的列联表，如 2×2 四格表。系数为 1 表示两者完全一致，系数为 0 表示两者没有关联。

④ **【风险】**。计算相对危险度和比数比。表明事件的发生和某因素之间的关联性，如检验心脏病是否与吸烟有关。如果该系数的置信区间包括 1，则认为事件的发生与这个因素没有关联。当某因素发生的可能性非常小时，使用比数比统计量作为相对危险度的测度。

⑤ **【McNemar】**。两个二分变量相关性的非参数检验。在“试验前后”的设计中，变化值符合卡方分布。对检验由于试验干扰而产生的变化十分有效。

⑥ **【名义】**栏。计算两个名义变量之间关联程度的统计量。

- **【相依系数】**。相依系数是描述两个分类变量之间关联程度的统计量，根据卡方统计量修正而得，公式为

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

式中， $N$  为观测量数； $\chi^2$  为卡方值。其数值在 0~1 之间。0 值表示行列变量之间没有关联；其值接近 1，表示行列变量之间有很强关联。注意，该值可能的最大值受列联表行列数的影响，随着行列数的增大而增大。

- **【Phi and Cramer 变量】**（此处的变量应为  $V$ ），同相依系数一样，也是描述两个属性变量间关联程度。Phi 系数适用 2×2 交叉表。二者都是根据卡方计算公式修正而得，Cramer  $V$  值在 0~1 之间，其计算公式为

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

式中， $k$  为行、列变量中水平数较小的一个水平数； $N$  为观测量数； $\chi^2$  为卡方值。

- **【Lambda】**。当用自变量预测因变量时，该检验反映预测误差。Lambda 系数等于 1，表明自变量完全预测因变量；Lambda 系数等于 0，表明自变量不能预测因变量。
- **【不定性系数】**。常称为不确定性系数，表示用一个变量预测另一个变量的值可能发生的错误程度。其值越接近其上限 1，表明从第一个变量获得的有关第二个变量的值的信息越多；越接近其下限 0，表明从第一个变量获得的有关第二个变量的值的信息越少。程序计算对称与不对称两种不确定性系数。

⑦ **【有序】**栏。计算两个有序变量之间关联程度的统计量。

- **【Gamma】**。两个有序变量间关联的对称检验，该值范围在-1~1 之间。Gamma 的绝对值接近 1，表明两个变量间高度关联；接近 0，表明两个变量间的关联程度很低。对二维列联表，提供零阶 Gamma 值；对三维或高维列联表，提供条件 Gamma 值。
- **【Somers'd】**。两个有序变量间关联性的检验，其数值范围为-1~1。Somers'd 的绝对值接近 1 时，表明两个变量间高度关联；接近 0，表明低度关联。Somers'd 检验是 Gamma 的非对称检验的扩展，两者之间的不同仅在于它包含的是未打结自变量成对数据的含量。
- **【Kendall 的 tau-b】**。秩变量或等级变量关联性的非参数检验，计算中考虑结的影响。值的范围-1~1，符号表明两者间关系的方向。绝对值表明相关程度，只有在正方形表格中其值才有可能为+1 或-1。
- **【Kendall 的 tau-c】**。秩变量关联性的非参数检验，不考虑结的影响。其值的范围是+1~-1，符号表明两者间关系的方向，绝对值表明相关程度。如果交叉表边际频数相等，那么 Kendall's tau-b 和 Kendall's tau-c 的值基本一致。

⑧ **【按区间标定】**(应为**【名义与等间隔变量】**, nominal by interval) 栏。计算 Eta 统计量。如果一个变量是名义变量，另一个变量是定量变量时，选择 Eta 统计量。Eta 值的范围在 0~1 之间，值为 0 表示行列变量间没有关联性，值越接近 1 关联程度越高。

⑨ **【Cochran's and Mantel-Haenszel 统计量】**。两个二分类变量间独立性检验的统计量。在此框中设置相对风险检验的零假设值，默认为 1。可以输入一个正数。

(8) 单击**【精确】**按钮，打开**【精确检验】**对话框，见图 7-20。

除了非参数检验与交叉表检验外，精确检验提供两种专门针对数据量小或不均衡表的检验方法，该检验对数据没有要求。检验的方法有 Fisher 精确检验和 Monte Carlo 法。由于精确检验的计算复杂，对大样本会耗费大量的计算机资源，因此在样本量少于 30 时，这是最好的方法。

① **【仅渐进法】**。显著概率值是基于渐近分布计算的统计量。一般情况下，如果其显著水平值小于 0.05，认为有显著性意义。

② **【Monte Carlo】**。该统计量是精确显著水平的无偏估计。Monte Carlo 方法不要求渐近分布的假设，可获得精确的显著水平值。

- **【置信水平】**框。输入 0.01~99.9 置信水平。
- **【样本数】**框。输入 1~1000000000 之间的样本量数值，用以计算 Monte Carlo 统计量。样本量越大，显著水平越可靠，但计算过程耗时也越多。
- **【精确】**，精确计算检验的概率。此值如果小于 0.05，则认为行、列变量间相互不独立。当有 20%以上单元格的期望数小于 5 时，适合使用该方法。

选中**【每个检验的时间限制为】**项，在参数框中输入 1~9999999999 间的值作为进行精确检验的最大运行时间。当计算条件受到限制时，常使用 Monte Carlo 精确检验法。

(9) 在主对话框中，单击**【单元格】**按钮，打开**【交叉表：单元显示】**对话框，如图 7-21 所示。

- ① **【计数】**栏。指定交叉表中显示的计数选项。

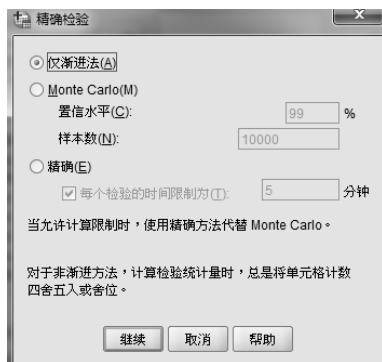


图 7-20 **【精确检验】**对话框

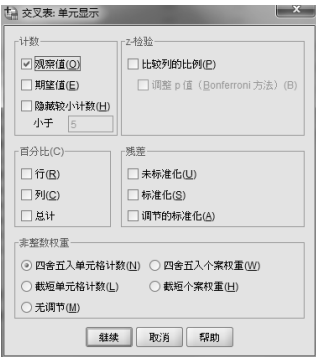


图 7-21 【交叉表：单元显示】对话框

- **【观察值】**。显示实际频数，这是默认选项。
- **【期望值】**。如果行、列变量在统计意义上相互独立，显示期望频数(理论数)。
- ② **【百分比】** 栏。指定输出的百分比。
- **【行】**。行百分比，单元格频数占所在行观测量的百分比。
- **【列】**。列百分比，单元格频数占所在列观测量的百分比。
- **【总计】**。单元格中频数占全部观测量的百分比。
- ③ **【残差】** 栏。指定要输出的残差。
- **【未标准化】**。计算非标准化残差。单元格中的观测值减期望值。

- **【标准化】**。均值为 0，标准差为 1 的标准化残差。残差除以它的标准误，也称为皮尔逊残差。
- **【调节的标准化】**。计算调整的标准化残差。
- ④ **【非整数权重】** 栏。选择非整数权重处理方法。单元格中的频数一般是整数，但是如果有带有小数的变量值加权，单元格的计数值可能出现小数。
- **【四舍五入单元格计数】**。照常使用观测量权重，但是单元格中累积权重需要在计算统计量之前四舍五入。
- **【截短单元格计数】**。照常使用观测量权重，但是单元格中累积权重需要在计算统计量之前截取整数部分。
- **【四舍五入个案权重】**。在加权计算之前对权重值四舍五入。
- **【截短个案权重】**。在加权之前对权重值截取整数部分。
- **【无调节】**。不对单元格数值进行调整。

(10) 单击**【格式】**按钮。打开**【交叉表：表格格式】**对话框，见图 7-22，确定表格中从左到右频数的排列顺序。

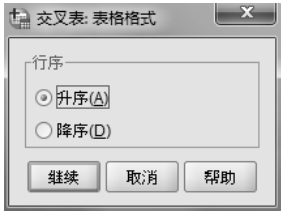


图 7-22 【交叉表：表格格式】对话框

- ① **【升序】**。以升序显示变量值的频数，这是默认选项。
- ② **【降序】**。以降序方式显示变量值的频数。

对长字符型变量，可以通过编码满足该过程对数据的要求。

7.4.3 交叉表分析实例

**【例 5】** 使用数据文件 data07-01 中的数据。使用变量：occcat80 工作性质分类、region 地区、childs 每个家庭的孩子数。要求分析各地区工作类型与家庭孩子数之间是否有关联。

- (1) 按**【分析→描述统计→交叉表】**顺序单击菜单项，打开主对话框。
- (2) 将**【孩子数量[childs]】**作为行变量送入**【行】**框中；将**【工作分类[occcat80]】**作为列变量选入**【列(C)】**框中；将**【地区分类[region]】**变量选入**【层 1 的 1】**框中，作为层变量。
- (3) 单击**【统计量】**按钮，打开**【统计量】**对话框，选中**【卡方】**选项。单击**【继续】**按钮，返回主对话框。
- (4) 单击**【单元格】**按钮，打开**【单元显示】**对话框，确认**【计数】**栏中选中**【观察值】**选项。单击**【继续】**按钮，返回主对话框。
- (5) 打开**【精确】**对话框，选择**【Monte Carlo】**，在**【样本数】**中输入样本数量“1517”。单击**【继续】**按钮，返回主对话框。

(6) 单击【格式】按钮，打开【交叉表：表格格式】对话框，选择【升序】选项。单击【继续】按钮，返回主对话框。

(7) 在主对话框中，单击【确定】按钮，提交执行。

(8) 输出结果见表 7-12～表 7-14，分析如下：

表 7-12 观测量统计处理摘要

	案例					
	有效的		缺失		合计	
	N	百分比	N	百分比	N	百分比
孩子数量*工作分类*地区分类	1414	93.2%	103	6.8%	1517	100.0%

表 7-12 所示为观测量处理摘要，包括 N(样本量)、百分比、缺失值。  
表 7-13 所示为不同地区、不同工作性质与不同家庭孩子数量的交叉表。  
表 7-14 所示是独立性的卡方检验结果。

- 注意：由于许多单元格的频数少于 5，所以应该用 Fisher 精确检验结果得出结论。
- ① 东北部地区，Fisher 精确检验 Monte Carlo 双侧显著性概率为 0.117，大于 0.05，所以家庭中拥有孩子的数量与工作类型没有关联。
- ② 东南部地区，Fisher 精确检验 Monte Carlo 双侧显著性概率为 0.011，小于 0.05，所以家庭中拥有孩子的数量与工作类型有关联。
- ③ 西部地区，Fisher 精确检验双侧显著性概率为 0.072，大于 0.05，所以结论与①相同，即家庭中拥有孩子的数量与工作类型没有关联。

表 7-13 各变量之间的多维交叉列联表

孩子数量*工作分量*地区分类交叉制表								
Count		工作分类						合计
地区分类		管理者或者专 业人员	技术人员、销 售、行政人员	维修人员	农林渔业	精密或手工制 造业	一般操作人员 或装配人员	
东北部	孩子数量 0	44	57	21	3	19	16	160
	1	27	45	12	3	10	18	115
	2	41	61	23	2	15	27	169
	3	21	43	14	0	8	14	100
	4	12	11	11	0	11	13	58
	5	1	6	3	1	2	6	19
	6	1	2	1	0	0	1	5
	7	3	2	2	1	1	3	12
	8或8以上	0	0	0	0	1	0	1
	合计	150	227	87	10	67	98	639
东南部	孩子数量 0	29	31	11	5	8	19	103
	1	22	22	4	4	6	11	69
	2	10	20	22	2	10	20	100
	3	6	11	5	2	10	6	40
	4	3	7	11	0	5	7	33
	5	4	7	0	0	2	5	18
	6	1	1	3	0	0	3	8
	7	0	1	3	0	0	2	6
	8或8以上	0	1	0	0	1	2	4
	合计	83	109	59	13	42	75	381
西部	孩子数量 0	37	45	12	5	17	13	129
	1	17	18	6	0	6	12	59
	2	25	29	17	2	10	8	91
	3	11	19	13	3	11	4	61
	4	5	8	2	3	5	4	27
	5	6	0	2	0	2	2	12
	6	2	1	4	0	1	1	9
	7	1	0	1	0	0	0	2
	8或8以上	2	0	1	0	1	0	4
	合计	106	120	58	13	53	44	394

【例 6】小样本的列联表分析实例。

数据文件 data07-04 为某公司经理收入情况数据。使用变量：sex 性别、earnings 收入高低。分析男、女经理的收入高低是否不同。数据中有 15 个经理，其中男性 9 人，女性 6 人。由于样本较小，又是 2×2 交叉表，所以使用 Fisher 精确检验的结果。

表 7-14 卡方检验结果

卡方检验										
地区分类		值	df	渐进 Sig. (双 侧)	Monte Carlo Sig.(双侧)			Monte Carlo Sig.(单侧)		
					Sig.	99% 置信区间		Sig.	99% 置信区间	
						下限	上限		下限	上限
东北部	Pearson 卡方	47.163 <sup>d</sup>	40	.203	.186 <sup>b</sup>	.160	.212			
	似然比	44.483	40	.289	.262 <sup>b</sup>	.233	.291			
	Fisher 的精确检验	48.225			.117 <sup>b</sup>	.095	.138			
	线性和线性组合	9.514 <sup>e</sup>	1	.002	.003 <sup>b</sup>	.000	.006	.002 <sup>b</sup>	.000	.005
	有效案例中的 N	639								
东南部	Pearson 卡方	61.974 <sup>f</sup>	40	.014	.016 <sup>b</sup>	.008	.025			
	似然比	65.957	40	.006	.009 <sup>b</sup>	.003	.016			
	Fisher 的精确检验	55.621			.011 <sup>b</sup>	.004	.018			
	线性和线性组合	9.398 <sup>g</sup>	1	.002	.003 <sup>b</sup>	.000	.006	.001 <sup>b</sup>	.000	.002
	有效案例中的 N	381								
西部	Pearson 卡方	47.883 <sup>h</sup>	40	.183	.191 <sup>b</sup>	.165	.216			
	似然比	52.035	40	.096	.115 <sup>b</sup>	.094	.136			
	Fisher 的精确检验	47.618			.072 <sup>b</sup>	.055	.089			
	线性和线性组合	.683 <sup>i</sup>	1	.408	.411 <sup>b</sup>	.378	.443	.200 <sup>b</sup>	.174	.227
	有效案例中的 N	394								
合计	Pearson 卡方	73.038 <sup>a</sup>	40	.001	.003 <sup>b</sup>	.000	.006			
	似然比	70.970	40	.002	.001 <sup>b</sup>	.000	.002			
	Fisher 的精确检验	70.181			.000 <sup>b</sup>	.000	.003			
	线性和线性组合	17.826 <sup>c</sup>	1	.000	.000 <sup>b</sup>	.000	.003	.000 <sup>b</sup>	.000	.003
	有效案例中的 N	1414								

a. 17 单元格(31.5%) 的期望计数少于 5。最小期望计数为 .23。  
b. 基于 1517 采样表，启动种子为 2000000。  
c. 标准化统计量是 4.222。  
d. 28 单元格(51.9%) 的期望计数少于 5。最小期望计数为 .02。  
e. 标准化统计量是 3.084。  
f. 30 单元格(55.6%) 的期望计数少于 5。最小期望计数为 .14。  
g. 标准化统计量是 3.066。  
h. 32 单元格(59.3%) 的期望计数少于 5。最小期望计数为 .07。  
i. 标准化统计量是 .827。

(1) 操作步骤。读取数据文件后：

- ① 按【分析→描述统计→交叉表】顺序打开主对话框。
- ② 将【性别[sex]】选入【行】框中，将【收入 [earnings]】选入【列】框中。
- ③ 单击【统计量】按钮，对开【统计量】对话框，选中【卡方】选项。
- ④ 在主对话框中，单击【确定】按钮，提交系统执行。

(2) 输出结果见表 7-15～表 7-17。

由于样本过小，所有单元格的期望频数小于 5，最小的期望频数值为 2.8，Fisher 精确检验计算的双尾概率为  $p = 0.041$ ，小于 0.05，故结论是不同性别经理的收入高低差异显著。

表 7-15 观测量处理摘要

	案例处理摘要					
	案例					
	有效的		缺失		合计	
	N	百分比	N	百分比	N	百分比
性别*收入	15	100.0%	0	0.0%	15	100.0%

表 7-16 交叉列联表

性别*收入交叉制表				
计数		收入		合计
		低	高	
性别	男性	2	7	9
	女性	5	1	6
合计		7	8	15

表 7-17 卡方检验

卡方检验					
	值	df	渐进 Sig. (双 侧)	精确 Sig. (双 侧)	精确 Sig. (单 侧)
Pearson 卡方	5.402 <sup>a</sup>	1	.020	.041	.035
连续校正 <sup>b</sup>	3.225	1	.073		
似然比	5.786	1	.016		
Fisher 的精确检验					
线性和线性组合	5.042	1	.025		
有效案例中的 N	15				

a. 4 单元格(100.0%) 的期望计数少于 5。最小期望计数为 2.80。

b. 仅对 2x2 表计算

7.5 比率分析

常常见到两个尺度类型变量间比值的分析问题。例如，企业主营业务收入在总收入中的比重；篮球比赛中 3 分球得分占总得分的百分比；财产保险业务保费收入占全部业务保费收入的比例；汽车功率与车重之比等，它们都是比率的概念。如果希望计算比率，并将比率作为一个变量进行描述统计分析，可以使用比率分析过程。

先介绍几个基本概念。

(1) 平均绝对离差 AAD

它是各比率值与中位数之差的绝对值之和除以样本量，即

$$AAD = \frac{\sum |R_i - M|}{N}$$

式中， $R_i$  是比率变量值； $M$  是比率变量的中位数； $N$  是样本量； $i=1\sim N$ 。

(2) 离散系数 COD

它是比率变量平均差与中位数的比值，描述的是比率变量的离散程度。其公式是

$$COD = \frac{|R_i - \bar{R}|}{NM}$$

(3) 相关价格微分 PRD

也称为递减指数，是比率均值与加权比率均值之比。

(4) 基于中位数的变异系数 COV

是对比率变量离散程度的描述，是比率变量的标准差与中位数的百分比，其公式为

$$COV = \frac{1}{M} \sqrt{\frac{(R_i - M)^2}{N}}$$

(5) 基于均数的变异系数 COV

与统计学中所讲的变异系数的概念相同，只是这里的变量是一个比率变量，它是比率变量的标准差与均数的百分比。

7.5.1 比率分析过程

- (1) 按【分析→描述统计→比率】顺序打开如图 7-23 所示的【比值统计量】主对话框。
- (2) 将计算比率的分子变量送入【分子】框。
- (3) 将计算比率的分母变量送入【分母】框。

(4) 如果需要进行分组分析，将分组变量送入【组变量】框。

(5) 单击【统计量】按钮，进入如图 7-24 所示的【比率统计量：统计量】对话框，选择要输出的统计量。

①【集中趋势】栏。

- 【中位数】。输出比率的中位数。
- 【均值】。输出比率的均值。
- 【权重均值】。计算比率的加权均值。该值是用分子的均值除以分母的均值。
- 【置信区间】。计算比率的均数、中位数、加权均值 95% 的置信区间。在【级别 (%)】(应为【置信水平】)框内输入大于等于 0、小于 100 的数值作为置信水平。

②【离散】栏。输出比率变量离散趋势指标。

- 【AAD】。平均绝对离差。



图 7-23 【比值统计量】主对话框



图 7-24 【比率统计量：统计量】对话框

- 【COD】。离散系数。
- 【PRD】。相关价格微分。
- 【中位数居中 COV】。基于中位数的变异系数。
- 【均值居中 COV】。基于均数的变异系数。
- 【标准差】。比率的标准差。
- 【最小值】、【最大值】。比率变量的最小值和最大值。
- 【范围】。输出比率变量的全距。

③【集中指数】栏。用来测度落在置信区间内的比率百分比集中系数。可以通过两种方式进行计算：

- 在【介于比例】栏的【低比例】框内，输入指定区间的下限值，在【高比例】框内输入指定区间的上限值，然后单击【添加】按钮，计算落在这个区间的百分比。
- 在【中位数百分比之内】栏内，计算落在距离中位数这个区间内的比率数占比率总数的百分比。在【中位数百分比】框内输入 0~100 之间的数值，单击【添加】按钮。区间下限为  $(1-0.01 \times \text{该值}) \times \text{中位数}$ ，区间上限为  $(1+0.01 \times \text{该值}) \times \text{中位数}$ 。

(6) 在主对话框中选择结果输出方式：

- 【显示结果】。只在输出窗口显示结果。
- 【将结果保存到外部文件】。将结果保存为外部文件。

- **【按组变量排序】。**指定按分组变量输出的结果的升降序。选择**【升序】**或**【降序】**按分组变量值的升序(或降序)输出结果。

7.5.2 比率分析实例

**【例 7】** data07-05 是美国某州估税员按现有资源价值评估地产价值的假设数据文件。调查数据为过去一年中在该州售出的房产。该数据记录了每处房产在该州的位置(town)、估税员自评估以来持续观察地产的时间(time)、房产的售价(saleval)以及最终估价(lastval)。为了帮助政府追踪房产销售状况,合理公正地制定房产税,对估价与售价比进行分析。

(1) 操作步骤

- ① 按**【分析→描述统计→比率】**顺序打开**【比值统计量】**主对话框。
- ② 选择**【房产最终估价[lastval]】**变量,将其送入**【分子】**框,作为分子变量。
- ③ 选择**【房产售价[saleval]】**变量,将其送入**【分母】**框,作为分母变量。
- ④ 选择**【房产所在镇的位置[town]】**变量,将其送入**【组变量】**框,作为分组变量。
- ⑤ 单击**【统计量】**按钮,进入**【比率统计量:统计量】**对话框。在**【集中趋势】**栏选中**【中位数】**,在**【离散】**栏中只保留 COV 选项。
- ⑥ 在**【集中指数】**栏的**【低比例】**框内输入“0.8”,在**【高比例】**框内输入“1.2”,单击**【添加】**按钮,将其送入框内;在**【中位数百分比之内】**栏内的**【中位数百分比(%)】**框内输入“20”,单击**【添加】**按钮,将其送入框内。单击**【继续】**按钮,返回主对话框。
- ⑦ 单击**【确定】**按钮,提交运行,结果见表 7-18、表 7-19。

(2) 输出结果及解释

表 7-18 所示是对样本数据的描述摘要,给出了各地区房产数量和所占百分比。

表 7-19 所示是房产最终估价与售价的比率(估售比)统计量表。第 1 列是房产所在位置。第 2 列是比率的中位数,通过各镇房产估价与售价比率中位数的比较,可以判断哪个位置的房产估售比变化最大。本例中,北部的中位数是 0.963,接近 1,估售比变化最小;相反,南部房产估售比变化最大。第 3 列是离散系数 COD,它是描述比率变异大小的指标,数值越大,变异越大。本例中,北部的 COD 是 0.070,最小,说明北部房产的估售比变异最小;而南部的 COD 是 0.199,变异最大。最后两列是集中指数 COC,其中第 4 列是估售比落在 0.8~1.2 之间的百分比,北部是 95.9%,只有 4.1%房产是需要政府办公室重点关注的;南部该值为 36.1%,有 63.9%房产需要重点关注。最后一列是估售比落在中位数两侧 20%区间的占该区所售房产的百分比,其值越大,表明变异越小。

表 7-18 对样本数据的描述摘要

案例处理摘要			
		计数	百分比
房产所在镇的位置	东部	177	17.7%
	中心	187	18.7%
	南部	205	20.5%
	北部	220	22.0%
	西部	211	21.1%
总数		1000	100.0%
排除的		0	
总计		1000	

表 7-19 房产最终估价与售价的比率统计量

房产最终估价/房产售价的比率统计量				
组	中值	离散系数	集中系数	
			百分比介于 0.8 和 1.2 之间(包含 0.8 和 1.2)	在中值的 20% 范围内(包括 20%)
东部	.867	.128	67.2%	78.5%
中心	.904	.118	75.9%	81.8%
南部	.747	.199	36.1%	58.5%
北部	.963	.070	95.9%	95.9%
西部	.816	.118	55.5%	84.8%
总数	.873	.141	66.3%	75.7%



## 7.6 P-P 图和 Q-Q 图

P-P 图和 Q-Q 图都是根据累计分布函数理论计算的,使用它们可以进行数据是何种分布的检验,常用于检验数据是否服从正态分布。如果图形中所有点都聚集在直线上,则说明变量分布服从于所要检验的分布。

### 7.6.1 P-P 图和 Q-Q 图分析过程

(1) 按【分析→描述统计→P-P 图(或 Q-Q 图)】的顺序,打开如图 7-25 所示的【P-P 图(或 Q-Q 图)】主对话框。Q-Q 图与 P-P 图的界面是一样的。这里以 P-P 图为例进行介绍。

(2) 将一个或多个需检验的数值型变量送入【变量】框,对每个变量生成 P-P 图。

(3) 【检验分布】栏,指定检验的概率分布。提供 13 种概率分布:【Beta(贝塔分布)】、【卡方分布】、【指数分布】、【Gamma(伽玛分布)】、【半正态分布】、【Laplace(拉普拉斯分布)】、【Logistic(Logistic 分布)】、【Lognormal(对数正态分布)】、【正态分布】、【Pareto(帕雷托分布)】、【Student't (即 T 分布)】、【Weibull(威布尔分布)】、【均匀分布】。

如果选择了【T 分布】,还需要在【df(自由度)】参数框中输入自由度。

(4) 在【分布参数】栏中选择分布参数,选中【从数据中估计】选项,系统自动从数据中推算分布的参数,否则要在参数框中自行指定。选择的分布不同,参数框也不同。

(5) 在【转换】栏内选择变量转换方式。

①【自然对数转换】。将原变量值转换成以  $e$  为底的自然数值。

②【标准值】。将原变量转换成平均值为 0 和方差为 1 的标准正态分布  $Z$  值。

③【差分】。通过计算变量中连续两个数据之差来转换原有变量。输入一个正整数确定差分度。

④【季节性差分】。计算时间序列中两个恒定间距的数据差,用来转换原有时间序列数据,数

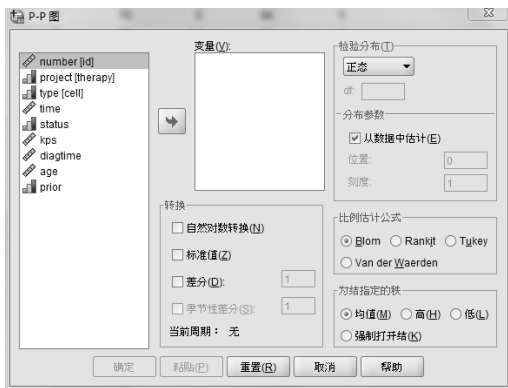


图 7-25 【P-P 图】对话框

据间隔的大小是根据当前所选择的周期而定。输入一个正整数以确定差分度。为了计算季节差分,必须确定含有周期因素的数据变量,如一年中的月份。

⑤【当前周期】。用来指明计算时间序列的季节差分。如果当前周期为 0,不能计算季节差。在【数据】菜单中【定义日期】命令可以建立具有周期性的变量。根据已存在的时间序列使用【转换】菜单中的【创建时间序列】命令,可以建立新的时间序列变量。

(6) 【比例估计公式】栏。提供了不同的方法来近似正态分布。每次只能选择其中一项。

(7) 【为结指定的秩】栏。一个变量中多个相同的值构成结。在本栏中可以选用以下不同的方式解决结点处观测值的秩。

①【均值】。用打结观测值的平均秩来作为它们的秩值。

②【高】。用打结观测值中最高的秩来作为它们的秩值。

③【低】。用打结观测值中最底的秩来作为它们的秩值。

④【强制打开结】。绘制每个结点处的观测量,忽视权重的影响。

7.6.2 P-P 图和 Q-Q 图分析实例

【例 8】打开数据文件 data07-06, 检验肺癌患者生存时间变量 time 是否服从 Weibull 分布。

(1) 操作过程

按【分析→描述统计→P-P 图】顺序, 打开【P-P 图(或 Q-Q 图)】主对话框。在左侧变量框中选中【time】变量, 送入【变量】框中。在【检验分布】栏内, 单击下拉按钮, 选中【Weibull】, 其他为默认。单击【确定】按钮, 提交运行。

输出结果见表 7-20~表 7-22 和图 7-26。

结果解释表 7-20 所示是模型描述表, 主要描述所作 P-P 图的变量名是时间, 未作变量转换, 检验的分布是 Weibull 分布, 估计参数是尺度(scale)参数和形状(shape)参数。

表 7-21 所示是个案处理摘要。该数据共有 137 个时间序列, 无缺失值。

表 7-22 所示是估计的 Weibull 分布参数。该分布尺度参数是 110.127, 形状参数是 0.937。

图 7-26(a)所示为肺癌生存时间的 Weibull 分布 P-P 图。可以看到, 各点都基本在直线上, 因此可以得出该数据的分布呈 Weibull 分布。

表 7-20 模型描述  
模型描述

模型名称	MOD_1
序列或顺序	1
转换	无
非季节性差分	0
季节性差分	0
季节性期间的长度	无周期性
标准化	未应用
分布	Weibull
类型	
标度	估计
形状	估计
部分排序估计方法	Blom
为结指定秩	同数的值的扶均值

正在应用来自 MOD\_1 的模型指定。

表 7-21 个案处理摘要  
个案处理摘要

	time
序列或顺序长度	137
图中的缺失值数	用户缺失 0
	系统缺失 0

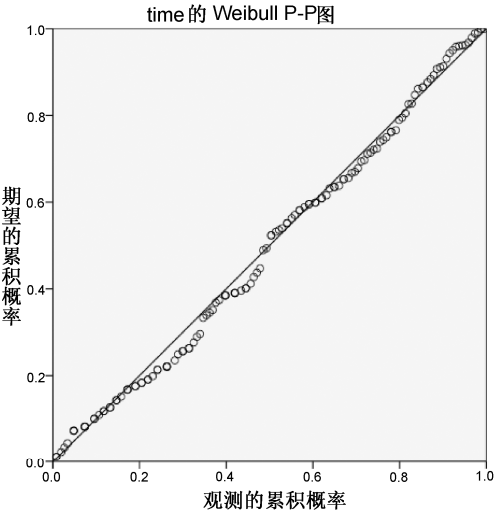
个案未进行加权。

表 7-22 估计的分布参数

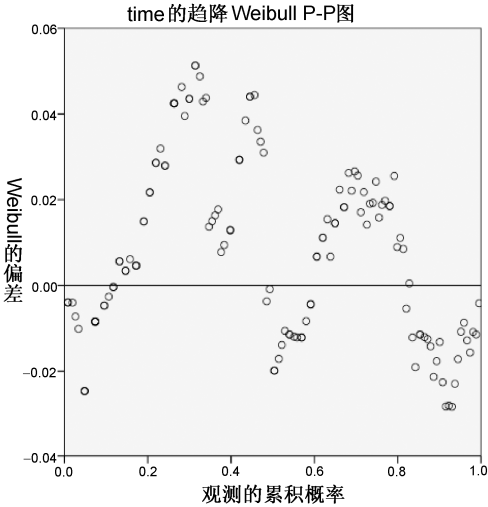
估计的分布参数

	time
Weibull 分布	标度 110.127
	形状 .937

个案未进行加权。



(a)



(b)

图 7-26 肺癌生存时间的 Weibull 分布 P-P 图和趋降 P-P 图

图 7-26 (b) 所示为肺癌生存时间的趋降 Weibull 分布 P-P 图。该图各点是无规则的，表明是随机的。

**【例 9】** 数据文件 data07-07 是 200 例正常人血铅含量数据文件，用 P-P 图分析过程检验 pb 变量是否服从正态分布。

操作步骤：按【分析→描述统计→P-P 图】顺序单击菜单项，打开【P-P 图(或 Q-Q Plots)】主对话框。将【血铅含量[pb]】变量选入【变量】框内，在【检验分布】框中选择【正态】，其他选项均为默认选项，单击【确定】按钮，提交系统运行。运行结果参见图 7-27。从图中看到，各点的分布没有完全在直线上，因此得出分布不呈正态分布。

现将 pb 变量的数据转换成自然对数数据，检验转换后的 pb 变量是否服从正态分布。  
操作步骤：操作与前面相同。只是在【转换】栏中选中【自然对数转换】选项，运行结果见图 7-28。

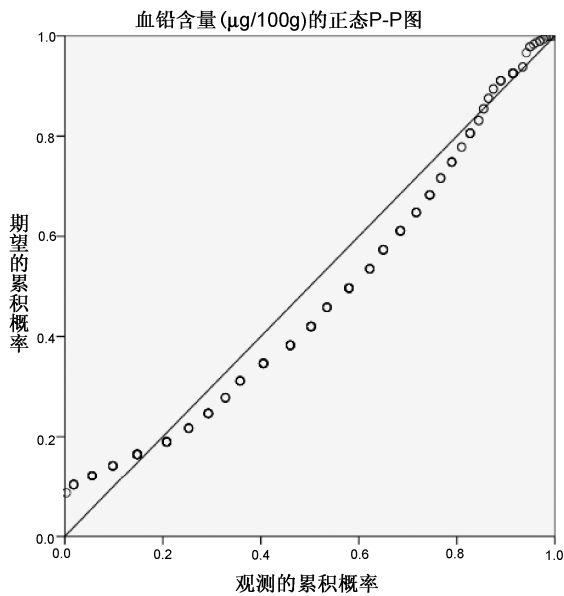


图 7-27 对 pb 变量正态性检验的 P-P 图

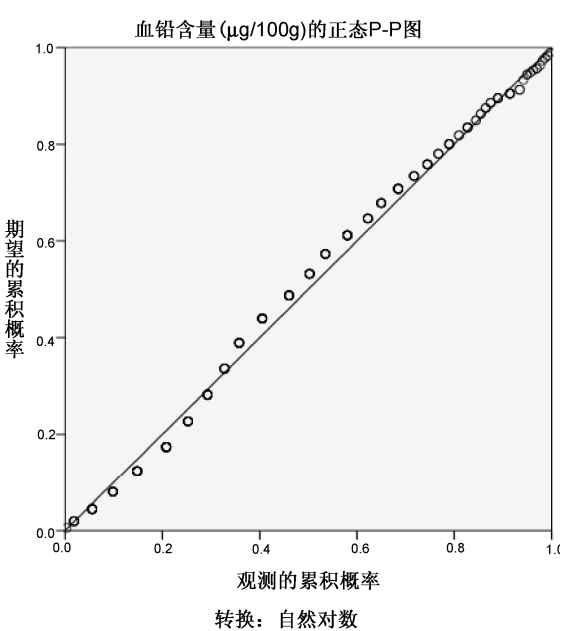


图 7-28 对转换后 pb 变量正态性检验的 P-P 图

从图 7-28 可以看出，所有点都在直线上，因此可以得出转换后的数据分布是近似正态的。结论是对非正态分布的变量，经过转换后呈正态分布，这时就可以用转换后的数据使用正态性假定的方法进行了。

**【例 10】** 数据文件 data07-08 是某市 150 名 3 岁女童身高的数据文件，使用 Q-Q 图分析过程检验身高数据的分布是否是正态分布。

操作过程：按【分析→描述统计→Q-Q 图】顺序，打开【Q-Q Plots 图】主对话框。将变量【身高[height]】选入【变量】框内，在【检验分布】框中选择【正态】选项，其他选项均为默认选项，单击【确定】按钮，提交系统运行。运行结果见图 7-29、图 7-30。

从图 7-29 可以看到，所有点都在直线上，因此可以得出 150 名 3 岁女童的身高数据分布呈正态分布。图 7-30 的各点是无规则的，表明是随机的。

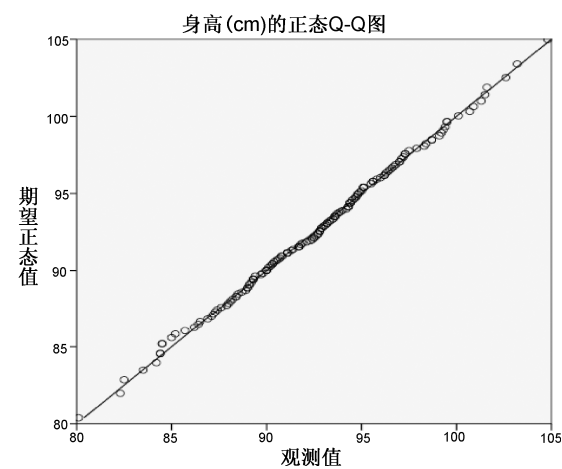


图 7-29 3 岁女孩身高的正态 Q-Q 图

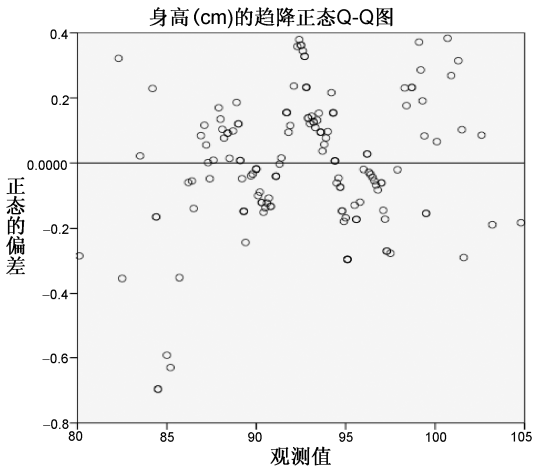


图 7-30 3 岁女孩身高的趋降正态 Q-Q 图

### 习 题 7

1. 对二维交叉表中两个变量间是否独立的检验，SPSS 提供了几种方法？各适合什么条件？单元格频数小于 5 时，应该考虑用什么方法检验？
2. 正态分布的变量用哪些描述统计量描述？哪些统计量描述该变量值的集中趋势？哪些统计量描述其离中趋势？
3. 如果变量的数据分布不是正态的，可以用哪些统计量描述？
4. 对数据明显为非正态分布的变量能用均值描述其平均水平吗？如果不能，使用什么指标描述比较合适？
5. 使用数据文件 data07-09，使用交叉表分析收入类型(inccat)与订阅报纸(news)之间的关系。
6. 使用数据文件 data07-09，利用频数表简单说明家庭收入(income)数据的分布情况。
7. 使用数据文件 data07-10，利用探索过程分析不同质量等级(标准、高级)与合金形成温度是否有关。

# 第 8 章 均值比较与检验

## 8.1 均值比较与均值比较的检验

### 8.1.1 均值比较的概念

统计分析常常采取抽样研究的方法，即从总体中随机抽取一定数量的样本进行研究来推论总体的特性。由于总体中的个体间存在差异，即使严格遵守随机抽样原则也会由于多抽到一些数值较大的或数值较小的个体致使样本统计量与总体参数之间有所不同。试验者测量技术的差别或测量仪器精确程度的差别等也会造成一定的偏差，使样本统计量与总体参数之间存在差异。由此可以得出这样的认识：均值不相等的两个样本不一定来自均值不同的总体。能否用样本均数估计总体均数，两个变量均值接近的样本是否来自均值相同的总体？换句话说，两个样本中某变量均值不同，其差异是否具有统计意义，能否说明总体差异？这是各种研究工作中经常提出的问题。这时就要进行均值比较。

### 8.1.2 均值比较与检验的过程

SPSS 提供以下计算变量描述统计量的过程和对均值进行检验的过程。这些过程用【分析】菜单中的【比较均值】菜单调用，见图 8-1。

#### 1. 均值过程的功能与术语

(1) 均值过程计算指定变量的综合描述统计量。

当观测按一个分类变量分组时，均值过程可以进行分组计算。例如，要计算工作人员上班路程的平均千米数，

SEX 变量把工作人员按性别分为女人、男人两组，均值过程可以分别计算男人、女人上班路程的千米数。用于形成分组的变量应该是其值数量少且能明确表明其特征的变量，可以是标称变量，如性别、民族、信仰等；也可以是顺序变量，即其值表明等级的，如年级、职称等。

使用均值过程求若干组的描述统计量，目的在于比较。因此必须使用分类变量，根据分类变量的值对因变量分组求均值。这是与描述统计过程不同之处。

(2) 均值过程中使用的术语

- ① 水平数。指分类变量的值的个数。例如性别变量有 2 个值，称为有两个水平。
- ② 单元。指因变量按分类变量值所分的组。例如，可以按性别将因变量的值分为两组。如果还有一个分类变量年龄，共有 10、11、12 三个值，可以将因变量分为 3 组。每组因变量的值称为一个单元。均值过程对每个单元的因变量值求各种描述统计量。
- ③ 水平组合。如果有两个分类变量，如性别(男、女)和年龄(10 岁、11 岁、12 岁)，按它

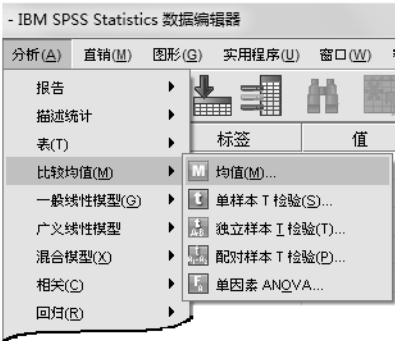


图 8-1 均值比较功能的调用

们的水平组合将会分因变量为 6 个单元,即男性 10 岁、男性 11 岁、男性 12 岁、女性 10 岁、女性 11 岁、女性 12 岁。

均值过程使用【比较均值】菜单中的【均值】项调用,见图 8-1。

## 2. T 检验过程

T 检验的过程按不同的比较方式分为 3 个功能:

### (1) 单一样本 T 检验

检验单个变量的均值是否与给定的常数之间存在差异。样本均数与总体均数之间的差异显著性检验属于单一样本 T 检验。例如,方便面面饼标准质量为 80g,可以看作总体均数。从生产线上任意抽取 100 个面饼,研究其平均质量与标准质量之间差异是否显著的问题就属于单一样本 T 检验。

### (2) 两个独立样本的 T 检验

两个独立样本的 T 检验用于检验两个不相关的样本来自具有相同均值的总体。例如,想知道购买某产品的顾客与不购买该产品的顾客平均收入是否相同,可以使用对两个独立样本进行 T 检验的功能。必须注意,使用这种检验的条件是必须具有来自两个不相关组的观测,其均值必须是对在两组中相同的变量的测度。

如果分组样本彼此不独立,如测度的是工人在技术培训前后某项技能的成绩,要求比较培训前后成绩均值是否有显著性差异,应该使用配对 T 检验的功能。如果分组不止两个,应该使用单因素 ANOVA 过程进行单变量方差分析。如果试图比较的变量明显不是正态分布的,则应该考虑使用一种非参数检验过程。如果试想比较的变量是分类变量,应该使用交叉表功能。

### (3) 配对样本 T 检验

配对样本 T 检验用于检验两个相关的样本是否来自具有相同均值的总体。这种相关的或配对的样本常常来自这样的试验结果,在试验中被观测对象在试验前后均被观测。例如,想要知道技术培训以后是否提高了工作效率,可以在培训前后测试完成一道工序的时间。在构成数据文件时,一个参与测试的工人在培训前后完成一道工序的时间形成一个观测,两个变量可以命名为 BEFORE 和 AFTER。配对分析的测度也不是必须来自同一个观测对象,一对可以两者组合而成。例如一对夫妻,或者试验前学习成绩和智商均相同的两个孩子作为一对,这样的若干对孩子分为两组,分别用不同教学方法进行教学,一段时间后,比较参与试验的两组学生平均成绩差异是否具有统计意义。在动物试验中常常把同一窝出生的体重、性别相同或最相近的小鼠配成试验的一对。

## 3. 单因素 ANOVA 过程

一元方差分析用于检验几个(3 个或 3 个以上)独立的组,是否来自均值相同的总体。例如,可以检验 3 个减肥训练计划,体重下降的效果(均值)是否相同,同时想看看哪一种训练计划效果最好,或者 3 个训练计划彼此之间哪两个之间的差异最显著,应该使用单因素 ANOVA 过程。如果按性别、体重级别对肥胖患者再进行分组,进行 3 个训练计划的试验。不但想知道哪一种训练计划对降体重下降最有效,而且想知道同一种训练方法对不同性别是否具有不同的效果,或者想去除每天的进食量对训练效果的影响,应该选择一般线性模型过程中子菜单中的各个功能进行多元方差分析或协方差分析。

如果分析变量明显是非正态分布的, 应该选择非参数检验过程, 非参数检验的内容请见第 12 章。单因素 ANOVA 过程以及多因素方差分析的内容请见第 9 章。

## 8.2 均值过程

均值过程的基本功能是分组计算, 比较指定变量的描述统计量, 包括均值、标准差、总和、观测数、方差等一系列单变量描述统计量, 还可以给出方差分析表和线性检验结果。

使用系统默认值即可按指定分组给出指定变量的均值、标准差、观测数等基本描述统计量。选项可以给出其他更加丰富的描述统计量。

### 8.2.1 均值过程中的统计量

如果变量为  $x$ ,  $x_i$  为变量  $x$  的第  $i$  个值, 共有非缺失观测数为  $n$  (或  $N$ ); 如果定义了加权变量  $w$ , 则  $w_i$  为第  $i$  个变量值对应的权重值, 可以选择的统计量关键字及含义如下:

(1) Sum 总和。加权和公式分别为

$$\text{Sum} = \sum_{i=1}^n x_i \quad \text{Sum} = \sum_{i=1}^n x_i w_i$$

(2) Number of Cases 观测数。如果定义了加权变量为  $w$ , 则公式为

$$N = \sum_{i=1}^n w_i$$

否则所有  $w_i = 1$ , 则  $N=n$ 。

(3) Mean 算术平均值。正态分布变量的集中趋势统计量, 公式为

$$\text{Mean} = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}$$

(4) Median 中位数。当变量值按大小排序,  $N$  为奇数, Median 是位置在正中间的值;  $N$  为偶数, Median 是两个位置在正中间的值之平均值。

(5) Grouped Median 分组中位数。每组变量值按大小排序,  $N$  为奇数, Median 是中值;  $N$  为偶数, Median 是两个中值之平均值。

(6) Variance 方差。正态分布变量的离散趋势统计量, 公式为

$$\text{Variance} = \frac{\sum_{i=1}^n w_i (x_i - \bar{x})^2}{\sum_{i=1}^n w_i - 1}$$

(7) Standard Deviation 标准差。公式为

$$S = \sqrt{\text{Variance}}$$

(8) Standard Error of Mean 均值的标准误。公式为

$$\text{Stderr} = \frac{S}{\sqrt{N}}$$

- (9) 最小值 Minimum。要求  $N \geq 1$ 。
- (10) 最大值 Maximum。要求  $N \geq 1$ 。
- (11) Range, 范围, 或称全距,  $\text{Range} = \text{Maximum} - \text{Minimum}$ 。
- (12) First 第一个变量值。按分组变量分组, 该组的第一个变量值。
- (13) Last 最后一个变量值。按分组变量分组, 该组最后一个变量值。
- (14) Kurtosis 峰度。是正态性检验统计量之一。其值为负, 分布曲线峰值高出正态分布曲线峰值; 其值为正, 分布曲线比较平坦。公式如下(要求  $N \geq 3, S > 0$ ):

$$\text{Kurtosis} = \frac{N^2 - 2N + 3}{(N-1)(N-2)(N-3)} \frac{\sum (x_i - \bar{x})^4}{S^4} - \frac{3(2N-3)}{N(N-1)(N-2)(N-3)} \frac{\left[ \sum (x_i - \bar{x})^2 \right]^2}{S^4}$$

- (15) Standard Error of Kurtosis 峰度的标准误。峰度的标准误。
- (16) Skewness 偏度。是正态性检验统计量之一。其值为正, 分布曲线相对于正态曲线左偏, 右尾较长; 其值为负, 分布曲线右偏, 左尾较长。公式如下(要求  $N \geq 2, S > 0$ ):

$$\text{Skewness} = \frac{N}{(N-1)(N-2)} \frac{\sum (x_i - \bar{x})^3}{S^3}$$

- (17) Standard Error of Skewness 偏度的标准误。
- (18) Percent of Total Sum 每组总和占整个观测总和的百分比。
- (19) Percent of Total N 每组中观测总数  $N$  占总观测数的百分比。
- (20) Geometric Mean 几何均数。主要用于变量值之间呈倍数关系的偏态分布, 公式为

$$G = \lg^{-1} \left( \frac{\sum \lg x_i}{N} \right)$$

- (21) Harmonic Mean 调和均数。主要用于求平均率、平均速度或平均存活时间等, 公式为

$$H = \frac{N}{\sum \frac{1}{x_i}}$$

8.2.2 均值过程操作

- (1) 建立的数据文件中要求至少有一个连续变量、一个分类变量(离散变量)。对连续变量求其基本描述统计量, 分类变量用来分组。



图 8-2 【均值】主对话框

- (2) 按【分析→比较均值→均值】顺序打开【均值】主对话框, 见图 8-2。
- (3) 选择因变量。在左面的变量表中选择要分析的变量作为因变量, 送入【因变量列表】框中。因变量可以选择一个, 也可以选择多个。
- (4) 自变量的选择及层控制。选择分组变量(也称自变量), 对因变量将按自变量分组计算基本描述统计量。选择的若干自变量可以放在第一层,

也可以放在其他层。

① 两个分类变量均放在第一层的操作是:



- 首先在源变量表中选择一个分类变量，送入【自变量列表】框中。此时层控制显示【层 1 的 1】，表示变量被送入第一层，建立了一个控制层。
- 在源变量表中选择第二个变量，送入【自变量列表】框中。此时层控制仍显示【层 1 的 1】，表示变量被送入第一层，共建立了一个控制层。该层有两个自变量。

例如，第一控制层的两个自变量分别有  $n_1$ 、 $n_2$  个水平，则程序运行结果，分别给出两个变量各水平的因变量的统计量，即按第一个自变量分  $n_1$  组给出因变量的描述统计量，按第二个自变量分  $n_2$  组给出因变量的描述统计量。

② 两个分类变量分别放在两层中的操作是：

- 在变量表中选择一个分类变量，送入【自变量列表】框中，建立了一个控制层。
- 单击【下一个】按钮，使层控制显示【层 2 的 2】，表明可以建立第二层了。
- 在变量表中选择第二个分类变量，将其送入第二层，显示在【自变量列表】框中作为第二层的分类变量。此时【上一个】、【下一个】两个按钮均加亮，表示既可以单击【上一个】按钮向前回到第一层，也可以单击【下一个】按钮，去建立第三层。

如果两个分类变量的水平数分别为  $n_1$ 、 $n_2$ ，并分别控制第一层和第二层，那么会将因变量分为  $n_1 \times n_2$  组，每个组合称为一个单元 (Cell)，按单元给出因变量的统计量。

综上所述，单元数的计算是同层变量的水平数相加，不同层的变量水平数相乘。

(5) 在主对话框中单击【选项】按钮，打开【均值：选项】对话框，见图 8-3。

① 【统计量】栏。选择统计量。

左面【统计量】栏内列出了可以计算的各组描述统计量，选择后单击向右箭头按钮将选定的统计量移至右面【单元格统计量】框中。可以选择的统计量关键字及含义如下：

合计、个案数、均值、中位数、组内中位数、方差、标准差、均值的标准误、最小值、最大值、范围、第一个 (按分组变量分组的该组的第一个变量值)、最后一个 (按分组变量分组，该组最后一个变量值)、峰度、峰度的标准误、偏度、偏度的标准误、总和的百分比 (每组总和占总和的百分比)、总个案数百分比 (每组中观测总数  $n$  占总观测数  $N$  的百分比)、几何均值、调和均值。

② 【第一层的统计量】栏。指定只对第一层每个控制变量进行的分析。

- 【Anova 表和 eta】。方差分析表和 eta 统计量  $\eta$ 、eta 平方统计量  $\eta^2$ 。方差分析检验的零假设是：第一层控制变量各水平上的因变量均值都相等。 $\eta$  统计量表明因变量和自变量之间联系的强度。 $\eta^2$  是因变量中不同组中差异所解释的方差比，是组间平方和与总平方和之比。
- 【线性相关检验】。产生平方和、自由度、均方、F 检验的  $F$  值、 $R$  和  $R^2$  等统计量。但在分类自变量是字符型时，不计算有关线性度的统计量。 $R$  和  $R^2$  是线性拟合的良好度的统计量，只有在控制变量有基本的数量级 (如控制变量表示年龄或药物剂量，不能是颜色

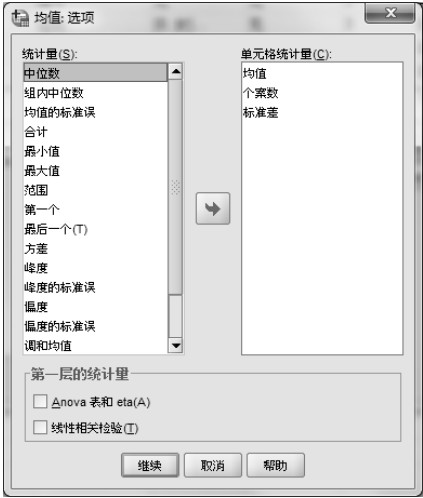


图 8-3 【均值：选项】对话框

或信仰等)，且自变量有三个水平以上时才计算。其假设的前提是因变量均值是第一层控制变量的线性函数。

8.2.3 分析实例

【例 1】数据文件 data08-01 是 27 名男女学生身高数据。数据文件中的变量顺序是：no 编号、sex 性别、age 年龄、h 身高、w 体重。要求按年龄分组比较身高均值；按性别分组比较身高均值。分析不同年龄和性别的学生身高均值。

(1) 对于不同年龄、不同性别学生的身高的分析，要把两个分类变量均放在第一层，操作如下：

- ① 按【分析→比较均值→均值】顺序打开【均值】主对话框，见图 8-2。
- ② 在源变量表中选择变量 h 作为因变量，送入【因变量列表】框中。
- ③ 在源变量表中选择分类变量性别 sex，送入【自变量列表】框中，再在源变量表中选择年龄变量 age，也送入【自变量列表】框中。建立了一个控制层，该层有两个分类变量。单击【确定】按钮，提交系统执行。
- ④ 运行结果见表 8-1 和表 8-2。

表 8-1 观测处理汇总表

	案例					
	已包含		已排除		总计	
	N	百分比	N	百分比	N	百分比
身高 * 性别	27	100.0%	0	0.0%	27	100.0%
身高 * 年龄	27	100.0%	0	0.0%	27	100.0%

表 8-2 基本描述统计量

身高			
性别	均值	N	标准差
女	1.5154	13	.06253
男	1.5357	14	.07623
总计	1.5259	27	.06941

(a)

身高			
年龄	均值	N	标准差
10	1.4488	8	.02167
11	1.5209	11	.03910
12	1.6129	7	.01704
13	1.5900	1	.
总计	1.5259	27	.06941

(b)

表 8-1 给出的是汇总信息。在第一控制层只给出两个分类变量：性别、年龄。因此进行的是(身高\*性别)按性别分组的身高均值比较和(身高\*年龄)按年龄分组的身高均值比较。该表“已包含”栏给出参与每个分析的观测数 N 均为 27 和占总观测数的百分比均为 100%；“已排除”栏给出每个分析中剔除的观测数，N 均为 0，占总观测数的百分比均为 0%；“总计”栏给出观测总数 N 和总百分比。

在 SPSS 统计分析过程执行的输出结果中，均给出这样的汇总表。在以后的章节中，如无特殊需要，不再进行解释或说明，或不再列出。

表 8-2 给出的是按性别和按年龄分组的分析结果。由于在定义系统参数时，要求输出显示变量标签和值标签，因此在表格中显示的分类变量不是原变量名和变量值，而是变量标签和值标签。

表 8-2(a) 分析变量是身高 h，分类变量是性别 sex。可以看出，女生 13 人平均身高为 1.5154，

标准差为 0.06253；男生 14 人平均身高为 1.5357，标准差为 0.07623；27 个学生总平均身高为 1.5259，标准差为 0.06941。

表 8-2 (b) 按年龄分组的结果是 10 岁 8 人平均身高为 1.4488，标准差为 0.02167；11 岁 11 人平均身高为 1.5209，标准差为 0.0391；12 岁 7 人平均身高为 1.6129，标准差为 0.01704；13 岁 1 人平均身高为 1.59，不能计算标准差，因此该项为缺失值。

(2) 另一种分析。发育阶段相同年龄的男孩和女孩是否身高有所不同？是否身高随年龄的增长呈线性关系？如果解决这样的问题，只建立一个控制层就不够了。应该考虑，选择身高 *h* 作为因变量，分类变量 *age* 作为第一层控制变量，*sex* 为第二层控制变量。两个分类变量分别放在两层中，且使用选项。操作如下：

① 按前面叙述的方法先将变量年龄 *age*，送入【自变量列表】框中建立一个控制层。单击【下一个】按钮，在变量表中选择第二个分类变量性别 *sex*，送入【自变量列表】框中，作为第二层。*age* 和 *sex* 分别控制第一层和第二层。

② 单击【选项】按钮，打开【均值：选项】对话框，见图 8-3。在第一层的【统计量】栏中选【Anova 表和 eta】和【线性相关检验】，单击【继续】按钮返回主对话框。

③ 在主对话框中单击【确定】按钮，则在输出窗中得到输出结果，见表 8-3~表 8-5。

④ 结果说明。观察输出结果，与上一种分析对比，可以看出层变量的作用，还可以看出使用系统默认统计量和使用选项对话框中确定输出的统计量之间的不同。

表 8-3 所示是由第一层变量 *age* 和第二层变量 *sex* 确定的各单元的身高均值。

表 8-4 所示是方差分析与线性度检验的结果，说明如下：

方差分析的变量信息是“身高\*年龄”，因变量 *h* 标签是“身高”，分组 BY 变量 *age* 标签为“年龄”。说明方差分析的要求是分析不同年龄的身高均值间是否存在显著性差异。

表 8-4 中各统计量的名称与各统计量之间的数学关系：

- 偏差平方和。
- 组间偏差平方和。从表中可以看出，组间偏差平方和 0.105 (显示值，非机内值)。它由两部分组成，线性 = 0.097，是由因变量与控制变量之间的线性关系引起的；线性偏差 = 0.008，不是由因变量与控制变量之间的线性关系引起的。
- 组内平方和 = 0.020 组内偏差平方和。各组内各观测相对于组均值的变异，有时也称其为误差变异。

表 8-3 各单元的身高均值表

身高		均值	N	标准差
年龄	性别			
10	女	1.4500	5	.02000
	男	1.4467	3	.02887
	总计	1.4488	8	.02167
11	女	1.5383	6	.02317
	男	1.5000	5	.04637
	总计	1.5209	11	.03910
12	女	1.6100	2	.01414
	男	1.6140	5	.01949
	总计	1.6129	7	.01704
13	男	1.5900	1	.
	总计	1.5900	1	.
总计	女	1.5154	13	.06253
	男	1.5357	14	.07623
	总计	1.5259	27	.06941

表 8-4 对第一层变量的方差分析结果

ANOVA 表					
			平方和	df	均方
身高 * 年龄	组间	(组合)	.105	3	.035
		线性	.097	1	.097
		线性偏差	.008	2	.004
	组内		.020	23	.001
	总计		.125	26	
					F
					显著性

- 总计偏差平方和。它等于组间偏差平方和 0.105 与组内偏差平方和 0.020 之和 0.125。df 是自由度，组间自由度为 3，组内自由度为 23。
- 均方。数值上等于偏差平方和除以自由度之商。
- F 值。数值上等于组间均方与组内均方之比值。从表中可以看出，组间偏差平方和为 0.105，自由度为 3，均方值为组间偏差平方和除以自由度，值为  $0.105/3 = 0.035$ ；组内偏差平方和为 0.020，自由度为 23，均方值为  $0.020/23 = 0.001$  (注意，因显示位数有限，此处是近似值)。
- 显著性。在四个年龄组学生身高均值相等的零假设下，获得各统计量的值或更极端值的概率。从表中可以看到，显著性概率近似为 0。组间均方远远大于组内均方，说明组间差异远远大于随机误差引起的组内差异。因此可得出结论：10 岁、11 岁、12 岁、13 岁学生的身高差异显著。

线性回归方程的偏差平方和为 0.097，均方为 0.097，F 值为 109.437。显著性值近似为 0.000，即小于 0.001，说明回归方程预测性能很好。这也可以用表 8-5 中 R 值为 0.879 接近 1 来说明。

表 8-5 所示为关联度的测度：

- Eta ( $\eta$ ) 值为 0.915，说明因变量与自变量之间联系紧密。Eta 方 ( $\eta^2$ ) 等于组间偏差平方和与总偏差平方和之比，即  $0.105/0.125 = 0.84$  (表中因各项计算误差导致为 0.838)。

表 8-5 关联度测度

相关性度量				
	R	R 方	Eta	Eta 方
身高 * 年龄	.879	.772	.915	.838

- Eta 是 0~1 之间的数，越接近 1，表明因变量(身高)与控制变量(年龄)关系越密切；如果 Eta = 0 则表明两个变量无关。
- R 是因变量 h 观测值与预测值之间的线性相关系数。虽然没有直接求出回归方程，但可以知道，R 值越接近 1 表明线性回归方程的预测性能越好，因变量与自变量之间的线性回归关系越好。
  - $R^2$  是线性模型的拟合良好度，有时称作确定系数，是在因变量中由回归模型解释的方差比例，其值的范围是 0~1。该值越小，表明模型对数据的拟合越不好。

## 8.3 单样本 T 检验

### 8.3.1 单样本 T 检验的概念

单样本 T 检验过程检验单个变量的均值是否与给定的常数之间存在差异。例如，研究人员可能想知道一组学生的 IQ 平均分数与 100 分的差异。

如果已知总体均数，进行样本均数与总体均数之间的差异显著性检验属于单样本的 T 检验。变量的样本均值为  $\bar{x}$ ，已知总体均值(或给定常数)为  $\mu_0$ ，检验的零假设是  $H_0: \bar{x} = \mu_0$ 。

计算公式为

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

式中， $s_{\bar{x}} = \frac{s}{\sqrt{n}}$  是均值的标准误；s 是变量的标准差。

单样本 T 检验过程对每个检验变量给出的统计量有：均值、标准差和均值的标准误。该过程计算每个数据值与总体均值之间差的平均值，进行该差值为 0 的 T 检验及计算该差值的置信区间，读者可以指定检验的显著性水平。

8.3.2 单样本 T 检验的实例

【例 2】 数据文件 data08-02 中的数据是 1973 年某市测量的 120 名 12 岁男孩的身高资料。已知该地区 12 岁男孩平均身高为 142.5cm，问该市男孩身高与该地区平均身高有否差异？

- (1) 建立无效假设  $H_0$ ，假设某市 12 岁男孩身高与该地区 12 岁男孩身高平均值相等。
- (2) 建立数据集，仅有一个变量 Height12 岁男孩身高。
- (3) 按【分析→比较均值→单样本 T 检验】顺序打开【单样本 T 检验】对话框，见图 8-4。
- (4) 在对话框中将唯一的变量【Height】从源变量栏移至【检验变量】框内。在【检验值】框中输入将该地区 12 岁男孩平均身高“142.5”，见图 8-4。
- (5) 单击【选项】按钮，打开【单样本 T 检验：选项】对话框，见图 8-5。在【置信区间百分比】栏选择系统默认值“95%”，【缺失值】栏选择系统默认的【按分析顺序排除个案】。单击【继续】按钮，返回主对话框。
- (6) 在主对话框中单击【确定】按钮，输出结果见表 8-6 和表 8-7。



图 8-4 【单样本 T 检验】对话框

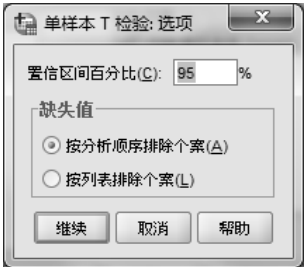


图 8-5 【单样本 T 检验：选项】对话框

(7) 结果分析。  
表 8-6 中的样本身高均值为 143.048，标准差为 5.821，标准误为 0.531。可以看出，样本均值 143.048 与地区身高平均值 142.5 比较，样本均值略高，差值为 0.548。

表 8-6 身高的基本描述统计量

单个样本统计量				
	N	均值	标准差	均值的标准误
12岁男孩身高	120	143.048	5.8206	.5313

表 8-7 中，t 值为 1.032，自由度为 119，双尾 T 检验的 P 值为 0.304>0.05，没有充分理由拒绝原假设。

表 8-7 单一样本 T 检验的分析结果

单个样本检验					
	检验值 = 142.5				
	t	df	Sig.(双侧)	均值差值	差分的 95% 置信区间
					下限    上限
12岁男孩身高	1.032	119	.304	.5483	- .504    1.600

当总体标准差未知时，差值的 95%置信区间=均值差值±1.96 × 标准误。根据表 8-7 得知，95%置信区间是 0.548 ± 1.96 × 0.531。由此推出，95%置信区间为 0.548 ± 196 × 0.531。这就是表 8-7 中“下限”与“上限”两项中的数值-0.504 和 1.600。这个 95%置信区间的含义是若以同样方式多次抽取等量的样本，对每个样本计算出的均值与总体均值的差异 95%落在这个区间之内。(注意：以上数值显示值与机内值有一定误差。因此如果使用计算器按显示值验算结果

会稍有不同，误差约小于 1%)。均值差值的 95%置信区间包括 0，没有充足理由拒绝样本均值与总体均值无显著差异的假设。

样本均值虽略高于总体均值，但无统计意义。误差来源可能是抽样误差，也可能来自测量误差。结论是，没有证据说明该市 12 岁男孩平均身高与该地区 12 岁男孩平均身高有显著性差异。

## 8.4 独立样本 T 检验

### 8.4.1 独立样本 T 检验的概念

进行独立样本的 T 检验要求被比较的两个样本彼此独立，即没有配对关系。要求两个样本均来自正态总体，要求均值是对于检验有意义的描述统计量。

两个样本方差相等与不等时使用的计算  $t$  值的公式不同。因此应该先对方差进行齐性检验。SPSS 的输出，在给出方差齐与不齐两种计算结果的  $t$  值，以及 T 检验的显著性概率的同时，还给出对方差齐性检验的  $F$  值和 F 检验的显著性概率。读者需要根据 F 检验的结果自行判断选择 T 检验输出中的哪个结果得出最后结论。

方差齐性检验的无效假设是：两个独立样本来自方差相等的两个总体  $v_1 = v_2$ ，进行 F 检验。 $F$  值计算公式为

$$F = \frac{\text{Max}(v_1, v_2)}{\text{Min}(v_1, v_2)}$$

式中， $v_1$ 、 $v_2$  分别为两个样本的方差。两个方差较大的一个除以两个方差中较小的一个，其比值为 F 检验的  $F$  值。

当  $p$  值小于 0.05 时，拒绝原假设，认为方差不齐，否则 ( $p$  值大于等于 0.05) 不足以在这个检验中拒绝原假设。(不排除在更多样本时或另一个检验方法时拒绝零假设。)

如果用  $\bar{x}_1$ 、 $\bar{x}_2$  表示两个样本的均值， $n_1$ 、 $n_2$  分别为两个样本的观测数目， $v_1$ 、 $v_2$  分别为两个样本的方差，方差齐 ( $v_1 = v_2$ ) 时与方差不齐 ( $v_1 \neq v_2$ ) 时计算  $t$  值使用的公式如下。

方差齐时公式为

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

式中，分母是两个样本均数之差的标准误，其中的  $S_c$  是合并方差，公式为

$$S_c = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

方差不齐时比较两个样本的均值，可以对变量进行适当的变换使样本方差具有齐性，再使用上述 T 检验计算公式进行计算与分析。SPSS 提供的函数可以实现对变量进行转换，也可用下述公式计算  $t$  值并进行检验，许多统计学书中称之为  $T'$  检验。SPSS 也在独立样本 T 检验过程的输出中提供方差不齐时计算  $t$  值的公式：

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{v_1}{n_1} + \frac{v_2}{n_2}}}$$

独立样本 T 检验与配对样本 T 检验均使用 T Test 过程，但调用该过程的菜单不同，对数据文件结构的要求和所使用的命令语句也有所区别。

8.4.2 独立样本 T 检验的过程

(1) 按【分析→比较均值→独立样本 T 检验】顺序，打开如图 8-6 所示的【独立样本 T 检验】主对话框。

- (2) 在源变量框中选择要进行检验的变量，将其送入【检验变量】框中。
- (3) 选择分组变量，将其送入【分组变量】框中，见图 8-6。
- (4) 单击【定义组】按钮，展开【定义组值】对话框，见图 8-6。

① 如果指定的分组变量是分类变量(测度类型为名义变量或顺序变量)，且只有两个值，单击【定义组】按钮打开如图 8-7(a)所示的对话框，在【组 1】和【组 2】框中分别输入作为第一组和第二组的分类变量值。

② 如果指定的分组是连续变量(测度类型为度量)，或者测度类型为名义或顺序但有多值，单击【定义组】按钮打开如图 8-7(b)所示的对话框。选择【使用指定值】，在【组 1】和【组 2】框中指定两个特定值，系统只对具有这两个值的因变量均值进行比较。



图 8-6 【独立样本 T 检验】主对话框



图 8-7 【定义组】对话框

选择【割点】选项，在【割点】后面的框中输入一个值，会将观测按分组变量值大于等于该值和小于该值分为两个组。检验在这两个组之间进行，比较其因变量在两组的均值间是否差异显著。

(5) 在主对话框中，单击【选项】按钮，打开【单样本 T 检验：选项】对话框，见图 8-5。

- ① 【置信区间百分比】。在该框中指定置信区间，系统默认值是 95%。可以在框中重新输入一个由读者指定的百分比值。
- ② 【缺失值】栏。选择对缺失值的处理方法。
  - 【按分析顺序排除个案】。带有缺失值的观测与分析有关时才被剔除。
  - 【按列表排除个案】。将检验变量、分组变量矩形框中指定的变量带有缺失值的观测剔除。

8.4.3 独立样本 T 检验的实例

【例 3】 以银行男女雇员当前工资为例，使用数据文件 data08-08，检验男女雇员当前工资是否有显著性差异。使用性别变量 gender 作为分类变量，比较当前工资 salary 变量的均值。

首先假设银行雇员工资服从正态分布。检验的原假设  $H_0$ ：不同性别雇员的当前工资均值相等，取  $\alpha = 0.05$ 。

- (1) 读取数据文件 data08-08，按【分析→比较均值→独立样本 T 检验】顺序，打开【独立样本 T 检验】主对话框，见图 8-6。按如下步骤操作，即可使用系统默认值进行检验。
- (2) 选择【salary】作为检验变量，单击上面一个箭头按钮，将其送入【检验变量】框中。
- (3) 选择【gender】变量作为分组变量，单击下面一个箭头按钮，将其送入【分组变量】框中，见图 8-6。
- (4) 单击【定义组】按钮，打开相应的对话框。在【组 1】框中输入“f”，为女雇员组；在【组 2】框中输入“m”，即男雇员作为第二组。
- 其余使用系统默认值。
- (5) 输出结果见表 8-8 和表 8-9。

表 8-8 分析变量的简单描述统计量

性别		N	均值	标准差	均值的标准误
当前工资	女	216	\$26,031.92	\$7,558.021	\$514.258
	男	258	\$41,441.78	\$19,499.214	\$1,213.968

表 8-8 中是分析变量的简单描述统计量。

左边第一栏为分析变量的标签“当前工资”和分类变量标签“性别”，下方是用值标签表示的分组变量值，分为两组，一组为“女”，另一组为“男”。N 给出各组观测数目，男为 258 人，女为 216 人；“均值”给出各组观测的分析变量均值。本例中男性雇员现平均工资为 \$ 41441.78，女性雇员现平均工资为\$26031.92。分组给出分析变量的标准差：男为\$19499.214，女为\$7558.021。男组均值标准误为 1213.97，女组为 514.26。

表 8-9 给出方差齐性检验结果，以及 T 检验和校正 T 检验两种方法，并分别计算出的检验结果。

表 8-9 独立样本 T 检验的结果

独立样本检验										
		方差方程的 Levene 检验		均值方程的 t 检验						
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	差分的 95% 置信区间	
									下限	上限
当前工资	假设方差相等	119.669	.000	-10.945	472	.000	-\$15,409.862	\$1,407.906	-\$18,176.401	-\$12,643.322
	假设方差不相等			-11.688	344.262	.000	-\$15,409.862	\$1,318.400	-\$18,002.996	-\$12,816.728

- ① 方差齐性检验(Levene 检验)结果，F 值为 119.669，显著性概率为  $p < 0.001$ ，因此结论是两组方差差异显著，即方差不齐。在下面的 T 检验结果中应该选择假设方差不相等一行的数据作为本例的 T 检验的结果数据。另一行是假设方差相等时 T 检验的计算结果，不取这个结果。
- ② “t” 栏显示两个值。本例的 t 值等于-11.69。“df” 栏给出两种 T 检验的自由度。
- ③ “Sig(双侧)” 是双尾 T 检验的显著性概率。本例的概率为 0.000，小于 0.05，否定不同性别雇员当前工资相等的原假设。可以得出结论男女雇员现工资具有显著差异。
- ④ 两组“均值差值”为-\$15409.9。平均现工资女雇员低于男雇员\$15409.9 元。
- ⑤ 差值的标准误为\$1318.40。
- ⑥ 差分的 95%置信区间在\$ - 18003.0~-12816.7 之间，不包括 0，也说明两组均值之差与 0 有显著差异。

结论：从 T 检验得  $p$  值为  $0.000 < 0.01$  和均值之差值的 95%置信区间不包括 0 都能得出，女雇员现工资明显低于男雇员，差异有统计意义。



注意：在实际应用中，由于存在其他条件，如职务等级、工作经验等，不能得出现平均工资差异是由性别差异造成的结论。根据分析结果得出结论要慎重。

【例 4】对连续变量按定点分组的独立样本 T 检验。

现对 data08-03 数据进行独立样本 T 检验。有 29 名 13 岁男生的身高、体重、肺活量数据，试分析身高大于等于 155 厘米与身高小于 155 厘米的两组男生的体重和肺活量均值是否有显著性差异。

首先建立无效假设  $H_0$ ：身高大于等于 155.0 与身高小于 155.0 两组之间的体重平均值在 99% 水平上无显著差异，两组之间的肺活量平均值在 99% 水平上无显著差异。

(1) 操作步骤

打开数据文件 data08-03，按【分析→比较均值→独立样本 T 检验】顺序，打开【独立样本 T 检验】主对话框。

在源变量表中选择体重 weight、肺活量 vcp 作为分析变量，用箭头按钮送入【检验变量】框中；选择变量身高 height 作为分组变量，用【箭头】按钮送入【分组变量】框中。

单击【定义组】按钮打开【定义分组】对话框。选择【割点】，并在框中输入“155.0”，见图 8-7(b)。单击【继续】按钮返回主对话框。

在主对话框中，单击【选项】按钮展开相应的对话框，见图 8-5。在【置信区间百分比】框中输入“99”，单击【继续】按钮返回主对话框。其他各选项使用系统默认值。单击【确定】按钮运行。

(2) 运行结果与分析

输出结果见表 8-10 和表 8-11。

表 8-10 分组描述统计量

身高		N	均值	标准差	均值的标准误
体重	>= 155.00	13	40.838	5.1169	1.4192
	< 155.00	16	34.113	3.8163	.9541
肺活量	>= 155.00	13	2.4038	.40232	.11158
	< 155.00	16	2.0156	.42297	.10574

表 8-11 方差齐性检验与 T 检验结果

		方差方程的 Levene 检验		均值方程的 t 检验						
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	差分的 99% 置信区间	
体重	假设方差相等	1.742	.198	4.056	27	.000	6.7260	1.6585	2.1309	11.3210
	假设方差不相等			3.933	21.745	.001	6.7260	1.7101	1.9004	11.5515
肺活量	假设方差相等	.002	.961	2.512	27	.018	.38822	.15456	-.04000	.81644
	假设方差不相等			2.525	26.277	.018	.38822	.15373	-.03859	.81504

表 8-10 给出了“体重”变量和“肺活量”变量按身高 height>=155.0 和 height<155.0 分组描述统计量。身高>=155.0 组 13 人，平均肺活量为 2.4038，平均体重为 40.838；身高<155.0 组 16 人，平均肺活量为 2.0156，平均体重为 34.113。

表 8-11 给出了方差齐次性检验和 T 检验的计算结果。从“Sig.(双侧)”栏数据可以看出，无论两组体重还是两组肺活量，方差均是齐的，均选择假设方差相等一行数据进行分析得出结论。

体重 T 检验的结果：Sig(双侧)=0.000，小于 0.001，当然小于 1%拒绝原假设。两组均值之差的 99%上、下限均为正值，也说明两组体重均值之差与 0 的差异显著。由此可以得出结论，按身高 155.0 分组的两组体重均值差异，在统计意义上高度显著。

肺活量 T 检验的结果: Sig(双侧)=0.018, 大于 0.01。两组均值之差的上、下限一个为正值, 一个为负值, 也说明差值的 99%上、下限与 0 的差异不显著。由此可以得出结论, 按身高 155.0 分组的两组肺活量均值差异在 99%水平上不显著。均值差异是由抽样误差引起的。

## 8.5 配对样本 T 检验

### 8.5.1 配对样本 T 检验的概念

进行配对样本的 T 检验要求被比较的两个样本有配对关系, 要求两个样本均来自正态总体, 要求均值是对于检验有意义的描述统计量。均值的配对比较是比较常见的, 例如:

(1) 同一窝试验用白鼠按性别、体重相同的配对, 再随机分到试验组和对照组, 分别喂加入海藻的饲料和普通饲料。3 个月后, 分别将每对白鼠置于水中, 测量其到溺死前的游泳时间, 比较两组白鼠游泳时间均值, 从而比较两种饲料对抗疲劳的作用。

(2) 同一组高血压病人, 在进行体育疗法前后, 测量其血压。每个病人在体育疗法前后的血压测量值构成观测对。可以求这组病人体育疗法前后的血压平均值, 进行配对 T 检验, 分析体育疗法对降血压的疗效。

(3) 研究人体各部位体温是否有差别, 一个人两个部位的温度构成一对数据。测量若干人同样两个部位的温度数据, 可以比较这两个部位的平均温度是否有显著性差异, 使用配对 T 检验。

配对样本 T 检验实际上是先求出每对测量值之差值, 再对差值变量求均值, 检验配对变量均值之间差异是否显著。其实质检验的假设, 是差值变量的均值与零均值之间差异的显著性。如果差值均值与 0 均值无显著性差异, 就说明配对变量均值之间无显著性差异。

如果差值变量为  $x$ , 差值变量的均值为  $\bar{x}$ , 样本的观测数为  $n$ , 差值变量的标准差为  $S$ , 差值变量的均值标准误为  $S_{\bar{x}}$ , 则配对样本 T 检验的  $t$  值计算公式为

$$t = \frac{\bar{x} - 0}{S_{\bar{x}}}, \quad S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

配对样本 T 检验与独立样本 T 检验均使用 T TEST 过程, 但调用该过程的菜单不同, 对数据文件结构的要求不同, 所使用的命令语句也有所区别。进行配对样本 T 检验的数据文件中, 一对数据必须作为同一个观测中的两个变量值。

### 8.5.2 配对样本 T 检验的过程

- (1) 建立数据文件。
- (2) 按【分析→比较均值→配对样本 T 检验】顺序, 打开【配对样本 T 检验】主对话框, 见图 8-8。
- (3) 指定配对变量。
  - ① 在主对话框的源变量表中选择一个变量, 单击向右箭头按钮, 变量名出现在【成对变量】框的【Variable1】列中。
  - ② 在源变量框中再选择一个与先选择的变量成对的变量, 单击向右箭头按钮, 变量名出现在【成对变量】框的【Variable2】列中。



图 8-8 【配对样本 T 检验】主对话框

可以使用上述方法指定多个配对变量。以上操作是使用系统默认参数进行配对样本 T 检验的基本操作。单击【确定】按钮就可以提交运行了。

(4) 配对样本 T 检验的选项

配对样本 T 检验使用系统默认值就可以得到比较满意的结果。如果想改变显著性概率，从而改变差值的置信区间，或者需要另外指定处理缺失值的方法，可以在单样本 T 检验主对话框中单击【选项】按钮，打开【单样本 T 检验：选项】对话框，见图 8-5，在对话框中改变系统默认值，指定需要的选项。有关操作参见 8.4.2 小节，这里不再赘述。

8.5.3 配对样本 T 检验的实例

【例 5】以体育疗法治疗高血压的数据为例，10 个高血压患者在施以体育疗法前后测定舒张压，数据文件为 data08-04。数据文件中的变量：number 编号、pretreat 治疗前舒张压(mmHg)、postreat 治疗后舒张压(mmHg)。要求判断体育疗法对降低血压是否有效。这是一个自身配对样本的 T 检验问题。解决问题的步骤如下：

首先建立无效假设( $H_0$ )：体育疗法对高血压病人舒张压的降低无疗效，即对高血压病人治疗前后舒张压的差值均数是由差值为 0 的总体中随机抽取的，差值不为 0 是由抽样误差引起的。

(1) 打开数据文件，按【分析→比较均值→配对样本 T 检验】顺序打开【配对样本 T 检验】主对话框，见图 8-8。

(2) 指定配对变量。配对变量为治疗前后的舒张压 pretreat 和治疗后舒张压 postreat。在主对话框左面的变量表中单击【pretreat】变量，按住 Ctrl 键并单击【postreat】变量，单击向右的箭头按钮，将配对变量送入【成对变量】框中。单击【继续】按钮，确认并返回主对话框，单击【确定】按钮提交运行。

(3) 运行结果见表 8-12～表 8-14。

表 8-12 治疗前后舒张压的简单描述统计量

	均值	N	标准差	均值的标准误
对 1 治疗前舒张压	119.50	10	10.069	3.184
治疗后舒张压	102.50	10	11.118	3.516

表 8-13 治疗前后舒张压相关系数

	N	相关系数	Sig.
对 1 治疗前舒张压 & 治疗后舒张压	10	.599	.067

表 8-14 对配对变量差值的 T 检验

	成对差分					t	df	Sig.(双侧)
	均值	标准差	均值的标准误	差分的 95% 置信区间				
				下限	上限			
对 1 治疗前舒张压 - 治疗后舒张压	17.000	9.534	3.015	10.180	23.820	5.639	9	.000

(4) 结果分析。

表 8-12 所示是对治疗前后舒张压的单变量描述统计量，表中显示的是配对变量的变量标签，对数为 1 对。

治疗前、后的舒张压均值分别为 119.50 与 102.50；N 即观测数目，治疗前后均为 10；治疗前、后的舒张压的标准差分别为 10.069 和 11.118；治疗前后的舒张压均值的标准误分别为 3.184 和 3.516。

表 8-13 中给出治疗前后舒张压之间的相关系数为 0.599, 不相关的概率为 0.067。相对于治疗前后舒张压的相关系数为 0 的假设成立的概率为 6.7%, 大于 5%, 可以得出结论: 治疗前后的舒张压没有明显的线性关系。

表 8-14 给出了配对变量差值的 T 检验结果。变量对均值之间的差值为 17.00; 差值的标准差为 9.53; 差值的标准误为 3.01; 差值的 95%置信区间上、下限分别为 10.18 和 23.82, 应注意两个值均为正值。

“t”值为 5.64; “df”自由度为 9; “Sig.(双侧)”是双尾 T 检验的结果, 获得 t 值的概率为 0.000, 即小于 0.001。

可以得出结论: 由于  $p$  小于 0.01, 拒绝原假设, 可以认为体育疗法对降低舒张压有明显疗效。

## 习 题 8

1. 均值比较的 T 检验分几种类型?

2. 要使用 T 检验进行均值比较的变量, 应该具有怎样的分布特征?

3. 两个独立样本 T 检验需要什么条件?

4. 一个品牌的方便面面饼的标称质量是 80 g, 不能大小相差很大, 因此要求标准差小于 2 g。现从生产线包装前的传送带上随机抽取部分面饼, 称重数据记录在数据文件 data08-05 中。问这批面饼质量是否符合要求?

5. 某康体中心的减肥班学员入班时的体重数据和减肥训练 1 个月后的体重数据记录在数据文件 data08-06 中, 试分析 1 个月的训练是否有效。如果按性别分组分析结果又如何? 如果按体重等级分组检查训练效果, 结果会是怎样的?

6. 为评价两个培训中心的教学质量, 对两个培训中心的学员进行了一次标准化考试, 考试成绩见数据文件 data08-07。分析两个培训中心教学质量是否有所差异, 得出统计分析结果, 并推断结论。

# 第9章 方差分析

## 9.1 方差分析的概念与方差分析过程

### 9.1.1 方差分析的概念

在科学试验中常常要探讨不同试验条件或处理方法对试验结果的影响，通常是比较不同试验条件下样本均值间差异。方差分析是检验多个样本均数间差异是否具有统计意义的一种方法。例如，医学界研究几种药物对某种疾病的疗效；体育科研中研究训练目标、方法和不同运动量等因素对提高某项运动成绩的效果；农业研究土壤、肥料、日照时间等因素对某种农作物产量的影响，不同饲料对牲畜体重增长的效果等，都可以使用方差分析方法去解决。

#### 1. 方差分析原理

方差分析的基本原理是认为不同处理组的均值间的差别基本来源有两个：

(1) 随机误差。例如，测量误差造成的差异或个体间的差异，称为组内差异，用变量在各组的均值与该组内变量值之偏差平方和的总和表示，记作  $SS_w$ ，组内自由度记作  $df_w$ 。

(2) 试验条件或不同的处理造成的差异，称为组间差异。用变量在各组的均值与总均值之偏差的总平方和表示，记作  $SS_b$ ，组间自由度记作  $df_b$ 。例如， $k \times m$  个试验对象随机分到  $k$  组，分别进行  $k$  种处理，要研究  $k$  种处理间均值是否存在显著差异，即处理是否有作用。测得数据是单因素  $k$  水平的完全随机设计数据，见表 9-1。

表 9-1 单因素  $k$  水平的完全随机设计数据

	$j=$ 处理 1	处理 2	处理 3	处理 4	...	处理 $k$
$i=1$	$x_{11}$	$x_{21}$	$x_{31}$	$x_{41}$	...	$x_{k1}$
2	$x_{12}$	$x_{22}$	$x_{32}$	$x_{42}$	...	$x_{k2}$
3	$x_{13}$	$x_{23}$	$x_{33}$	$x_{43}$	...	$x_{k3}$
4	$x_{14}$	$x_{24}$	$x_{34}$	$x_{44}$	...	$x_{k4}$
5	$x_{15}$	$x_{25}$	$x_{35}$	$x_{45}$	...	$x_{k5}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m$	$x_{1m}$	$x_{2m}$	$x_{3m}$	$x_{4m}$	...	$x_{km}$

其中， $i=1 \sim m$ ，是试验序号； $j=1 \sim k$ ，是处理序号； $x_{ij}$  是对第  $i$  个试验对象第  $j$  种处理后的因变量测试值。

此为平衡设计，即各处理组试验对象数相等，均为  $m$  个。数据的完全随机分析中，可以证明，总偏差平方和分解为组间偏差平方和和组内偏差平方和之和，即  $SS_t = SS_b + SS_w$ 。

总均值计算公式为

$$\bar{\bar{x}} = \frac{\sum_{j=1}^k \sum_{i=1}^m x_{ij}}{km}$$

第  $j$  种处理组均值为

$$\bar{x}_j = \frac{\sum_{i=1}^m x_{ij}}{m}$$

总偏差平方和如下:

$$SS_t = \sum_{j=1}^k \sum_{i=1}^m (x_{ij} - \bar{\bar{x}})^2$$

式中,  $x_{ij}$  是第  $j$  种处理组对第  $i$  个试验对象的观察值。

组间偏差平方和如下:

$$SS_b = m \sum_{j=1}^k (\bar{x}_j - \bar{\bar{x}})^2$$

反映处理间差异, 自由度  $df_b = k - 1$ 。

组内偏差平方和如下:

$$SS_w = \sum_{j=1}^k \sum_{i=1}^m (x_{ij} - \bar{x}_j)^2$$

总误差偏差平方和, 自由度  $df_w = k(m - 1)$ 。

为去除样本量的影响,  $SS_b$ 、 $SS_w$  除以各自的自由度得到其均方(即方差)值, 即组间均方和组内均方

$$MS_b = \frac{SS_b}{df_b}, \quad MS_w = \frac{SS_w}{df_w}$$

两者比值服从自由度为  $(k - 1)$  和  $k(m - 1)$  的 F 分布, 有

$$F = \frac{MS_b}{MS_w}$$

一种情况是处理没有作用, 即各样本均来自同一总体, 即  $MS_b/MS_w = 1$ ; 考虑抽样误差的存在, 则有  $MS_b/MS_w \approx 1$ 。

另一种情况是处理确实有作用, 组间均方是由于误差与不同处理共同导致的结果, 即各样本来自不同总体。那么, 组间均方会远远大于组内均方, 即  $MS_b \gg MS_w$ 。

$MS_b/MS_w$  比值构成 F 分布。用 F 值与其临界值比较, 推断各样本是否来自相同的总体。

## 2. 方差分析的假定条件和假设检验

### (1) 方差分析的假定条件

- ① 各处理条件下的样本是随机的。
- ② 各处理条件下的样本是相互独立的, 否则可能出现无法解释的输出结果。
- ③ 各处理条件下的样本分别来自正态分布总体  $N(\mu_i, \sigma_i^2)$ , 否则使用非参数分析。
- ④ 各处理条件下的样本方差相同, 具有齐性, 即  $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_k^2$ 。

### (2) 方差分析的假设检验

假设有  $k$  个样本, 如果原假设  $H_0$ : 样本均数都相同, 即  $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k = \mu$ ,  $k$  个样本有共同的方差  $\sigma^2$ , 则  $k$  个样本来自具有共同方差  $\sigma^2$  和相同均数  $\mu$  的总体。

如果经过计算, 组间均方远远大于组内均方, 使得  $F > F_{0.05(df_b, df_w)}$ , 则  $p < 0.05$ , 拒绝原假设, 说明样本来自不同的正态总体, 处理造成均值的差异有统计意义; 若  $F < F_{0.05(df_b, df_w)}$ , 则  $p > 0.05$ , 不拒绝原假设, 从而认为样本来自相同总体, 处理间无差异。

9.1.2 方差分析中的术语

方差分析中常用的术语如下。

1. 因素与处理

因素是影响因变量变化的客观条件，处理是影响因变量变化的人为条件，也可以统称为因素，实际上就是变量。例如，影响农作物产量的气温、降雨量、日照时间等为因素，研究不同肥料对不同种系农作物产量的影响时农作物的不同种系可称为因素，所施肥料可视为不同的处理。一般情况下，“因素”与“处理”在方差分析中可作相同理解。在要求进行方差分析的数据文件中均作为分类变量出现，即它们只有有限个取值。即使是气温、降雨量等平常看作连续变量的，在方差分析中如果作为影响产量的因素进行研究，也应该将其数值用分组定义水平的方法事先变为具有有限个取值的离散变量。

2. 水平

因素的不同等级称作水平。例如，性别因素在一般情况下只研究两个水平：男、女；化学或生物试验中的“剂量”必须离散化为几个有限的水平数，如 1 ml、2 ml、4 ml 三个水平。

3. 单元

在方差分析中，单元 Cell 是指各因素的水平之间的每个组合。例如，研究问题中的因素有性别 Sex，取值为 0、1；有年龄，分 3 个水平 1(10 岁)、2(11 岁)、3(12 岁)。两个变量的组合共可形成 6 个单元：[1,1]、[1,2]、[1,3]、[2,1]、[2,2]、[2,3]，代表两种性别与 3 种年龄的 6 种组合。在方差分析中，比较各单元条件下，因变量均值之间的差异。

4. 因素的主效应和因素间的交互效应

单独效应是在其他因素固定在某一水平时，因变量在某一因素不同水平间的差异。因素的主效应就是因变量在一个因素各水平间的平均差异。

当一个因素的单独效应随另一个因素的变化而变化时，称两个因素间存在交互效应。

这是在科学试验和生产实践中常常遇到的问题。举例说明，有 A、B 两种药物治疗缺铁性贫血，患者 12 例，分为 4 组。试验方案是：第一组用一般疗法，第二组在一般疗法基础上加用 A 药，第三组在一般疗法基础上加用 B 药，第四组在一般疗法基础上 A、B 两种药同时使用，1 个月

后观察红细胞增加数，分析两种药物的疗效，数据见表 9-2。

这是一个双因素方差分析的问题，因素 A 与因素 B 均有用该药与不用该药两个水平。研究药物 A 和 B 是否对红细胞的增加有显著影响需对红细胞增加数的均值作以下比较：

- ① 比较第二组的均值与第一组的均值是否有显著性差异。
- ② 比较第三组的均值与第一组的均值是否有显著性差异。

这两项研究的是 A、B 两因素的主效应。

除了比较第四组的均值与第一组的均值是否有显著性差异外，还要研究 A 药对 B 药的疗效是否有影响。若 A 药对 B 药疗效无影响，那么除采样误差外，第四组与第二组均值之差异

表 9-2 实验数据

	第一组	第二组	第三组	第四组
	红细胞增加数(×10 <sup>6</sup> /m <sup>3</sup> )			
	0.8	1.3	0.9	2.1
	0.9	1.2	1.1	2.2
	0.7	1.1	1.0	2.0
各组平均值	0.8	1.2	1.0	2.1

注：数据来源于《医用统计方法》(金丕焕，人民卫生出版社)。

该等于第三组均值减去第一组均值。但是实际上,  $2.1 - 1.2 = 0.9$ ,  $1.0 - 0.8 = 0.2$ , 相差 0.7, 该差值几乎与第一组均值相同。可以分析这个差异有统计意义, 0.7 的差值包括采样误差和 A、B 药的相互作用。这种因素之间的相互作用在统计学上称为交互效应, 在医学中称为协同效应(一个因素的单独效应随另一个因素的效应的增大而增大)或拮抗效应(一个因素的单独效应随另一个因素的效应的增大而减小)。如果交互效应存在, 说明两个因素不是相互独立的。

## 5. 均值比较

均值的相对比较是比较各因素对因变量的效应大小的相对比较。例如, 研究 A、B 效应之和是否等于它们的交互效应, 或者研究 A、B 药物对红细胞增加数的效应是否相等。

均值的多重比较是研究因素单元对因变量的影响之间是否存在显著性差异, 如上例中研究 A、B 药物对红细胞增加数的疗效是否存在显著性差异。

## 6. 单元均值、边际均值

在多因素方差分析中, 每种因素水平组合的因变量均值称为单元均值。一个因素水平的因变量均值称为边际均值, 这是根据它们在表格中的位置命名的, 参见 9.3.4 节中  $2 \times 2$  析因方差分析例题中的解释。

## 7. 协方差分析

在一般进行方差分析时, 要求除研究的因素外应该保证其他条件的一致。做动物试验往往采用同一胎的动物分组给予不同的处理, 研究不同处理对研究对象的影响就是这个道理。例如, 研究身高与体重的关系时要求按性别分别进行分析, 这样消除性别因素的影响。不同年龄的身高与体重的关系也是有区别的, 被测对象往往是不同年龄的。要消除年龄的影响, 应该采用协方差分析。再如, 研究几种饲料对增加动物体重的作用, 以便比较哪种饲料更好, 每个动物的进食量的影响应该在分析时消除, 也需要进行协方差分析。

## 8. 重复测量

组内变异的主要原因是实验对象之间的个体差异。由于个体差异存在, 即使试验对象受到相同的处理, 它们的因变量值也可能相对不同。重复测量设计的方差分析也是像协方差分析一样, 是在研究中减小个体差异带来的误差方差的一种有效方法, 而且由于对相同个体进行重复测量在一定程度上降低了人力、物力、财力的消耗。

如果重复测量是在一段时间内或一个温度间隔内进行的, 还可以研究因变量对时间、温度等自变量的变化趋势。这种重复测量研究称为趋势研究。例如, 将同一批动物在不同温度下生活一定时间并进行体重、脂肪的测定, 可以研究时间、温度对动物体重、脂肪量的变化趋势的影响。

### 9.1.3 方差分析过程

SPSS 提供的方差分析过程包括以下几种。

#### 1. 单因素 ANOVA 过程

单因素 ANOVA 过程是单因素的简单方差分析过程。它在【分析】菜单的【比较均值】过程组中, 见图 9-1, 用【单因素 ANOVA】菜单项调用。可以进行单因素的方差分析, 在方差相等或不相等的情况下进行均值多重比较和详细的对比。



## 2. 一般线性模型(General Linear Model, 简称 GLM) 过程

【GLM】过程由【分析】菜单直接调用。这些过程可以完成简单的多因素方差分析和协方差分析,不但可以分析各因素的主效应,还可以分析各因素间的交互效应。该过程允许指定最高阶次的交互效应,建立包括所有效应的模型。如果想建立包括某些特定的交互效应的模型,也可以通过【模型】对话框中的选项实现。均值多重比较、绘制轮廓图等功能对比较各因素各水平的单元格均值,直观地判断因素间的交互效应非常有用。

在【一般线性模型】菜单项的下一级菜单中有 4 项,见图 9-2。每个菜单项分别完成不同类型的方差分析任务。



图 9-1 单因素方差分析的菜单图



图 9-2 高级多元方差分析的菜单

### (1) 【单变量】过程

该过程提供回归分析和一个因变量与一个或几个因素变量的方差分析。因素变量把总体分为几组,可以检验关于其他变量在单个因变量各组均值效应的零假设,可以研究因素间交互效应以及单个因素(也可以是随机的因素)的效应。另外,还可以包括协变量效应和协变量与因素的交互效应。对回归分析,协变量指定作为自变量(预测变量)。在指定模型方面有较大的灵活性并可以提供大量的统计输出。

例如,如果以公司 4 个部门中的两个级别的职工为观察对象,研究生产率刺激机制。可以设计一个因子试验以便检验感兴趣的假设。由于在新刺激引入之前的原生产率可能对新刺激引入之后的生产率的比较产生很大影响,故可以把原生产率作为协变量进行协方差分析。如果想看协变量效应对两个级别的职工来说是否相同,也可以使用【单变量】菜单项调用【GLM】过程进行分析。

### (2) 【多变量】过程

该过程进行多因变量的多因素分析。当研究的问题具有两个或两个以上相关的因变量,要研究一个或几个因素变量与因变量集之间的关系时,才可以选用多变量菜单项。例如,研究数学、物理的考试成绩是否与教学方法、学生性别,以及方法与性别的交互作用有关时,使用此菜单项。如果只有几个不相关的因变量或只有一个因变量,应该使用【单变量】菜单项调用【GLM】过程。

多因变量的多因素分析过程同样可以研究因素间交互效应以及单个因素的效应,该因素可以是随机的因素。另外,还可以包括协变量效应和协变量与因素的交互效应。对回归分析,自变量(预测变量)指定为协变量,可以检验平衡和不平衡模型。

### (3) 【重复度量】过程

该过程进行重复测量方差分析。当一个因变量在同一课题中在不止一种条件下进行测量,

要检验有关因变量均值的假设时，应该使用该过程。如果指定了被试间因素，它们把总体划分成几个组，可以检验组间因素的效应和组内因素的效应的零假设，可以检验单个因素的效应以及因素间的交互效应，还包括协变量效应以及被试间因素与协变量之间的交互效应。

(4) 【方差分量估计】过程

该过程进行方差成分分析。通过计算方差估计值，可以帮助分析如何建立正确的模型。

9.2 单因素方差分析

单因素方差分析也称作一维方差分析。它检验由单一因素影响的一个(或几个相互独立的)因变量，由因素各水平分组的均值之间的差异是否具有统计意义，并可以进行两两组间均值的比较，称作组间均值的多重比较，还可以对该因素的若干水平分组中哪些组均值间不具有显著性差异进行分析，即一致性子集检验。

单因素 ANOVA 过程要求因变量属于正态分布总体。如果因变量的分布明显呈非正态，则不能使用该过程，而应该使用非参分析过程。如果对被观测对象的试验不是随机分组的，而是进行的重复测量形成几个彼此不独立的变量，应该用【重复度量】菜单项调用【GLM】过程对各因变量进行重复测量方差分析，条件满足时，还可以进行趋势分析。

9.2.1 简单的一维方差分析

【例 1】 用 4 种饲料喂猪，共 19 头猪分为 4 组，每组用 1 种饲料。一段时间后称重，猪的体重增加数据见表 9-3。比较 4 种饲料对猪体重增加的作用有无不同。数据文件为 data09-01。

1) 操作方法与步骤

(1) 在数据窗中建立数据文件，定义两个变量，并输入数据，这两个变量如下：

① fodder 变量，数值型，取值 1、2、3、4，分别代表 A、B、C、D 这 4 种饲料。

② weight 变量，数值型，其值为猪体重的增加数。

应该特别注意，不能把 A、B、C、D 定义为 4 个变量。

(2) 按【分析→比较均值→单因素 ANOVA】顺序单击菜单项，打开【单因素方差分析】主对话框，见图 9-3。

表 9-3 饲料比较数据资料

饲 料			
A	B	C	D
133.8	151.2	193.4	225.8
125.3	149.0	185.3	224.6
143.1	162.7	182.8	220.4
128.9	143.8	188.5	212.3
135.7	153.5	198.6	

注：数据来源于《医用统计方法》(金丕焕，人民卫生出版社)。



图 9-3 【单因素方差分析】主对话框

(3) 根据分析要求指定方差分析的因变量和因素变量。

① 选定 weight 变量进入【因变量列表】框中，定义猪体重增加数为因变量。

② 选定 fodder 变量进入【因子】框中，定义饲料为因素变量。

(4) 在主对话框中单击【确定】按钮，提交系统执行。

2) 输出结果(见表 9-4)

表 9-4 所示为因素变量饲料 fodder 对猪体重 weight 的影响分析结果。表的左上方是因变量的变量标签“猪体重增加量”。

(1) 输出结果说明

第一栏：方差来源。包括组间偏差 (Between Groups)、组内偏差 (Within Groups) 和偏差总和 (Total)。

第二栏：(偏差)平方和。组间偏差平方和为 20538.698，组内偏差平方和为 652.159，总偏差平方和为 21190.858，是组间偏差平方和与组内偏差平方和之和。

第三栏：自由度 df。组间自由度为 3，组内自由度为 15，总自由度为 18。

第四栏：均方。是第二栏与第三栏之比即偏差平方和除以自由度的结果。组间均方为 6846.233，组内均方为 43.477。

第五栏：F 值。是组间均方与组内均方之比。

第六栏：F 值对应的概率值。针对假设  $H_0$ ：组间均值无显著性差异(即 4 种饲料对猪体重增加的平均值无显著性差异)。计算的 F 值为 157.467，对应的概率值小于 0.001。

(2) 结果分析

根据输出的  $p$  值小于 0.001 可以看出，无论显著性水平取 0.05，还是取 0.01， $p$  值均小于临界值，因此拒绝  $H_0$  假设，认为 4 种饲料对猪体重增加的均值差异显著。结论：4 种饲料对猪体重的增加明显作用不同。

(3) 存在问题与解决方法

① 本例只考虑了猪体重的增加量，对其均值进行了比较，但实际工作中的问题往往不是这样简单的。例如，是否应该考虑每头猪的进食量对体重增加的影响，去除这个影响比较猪体重的增加会对饲料比较得出更切合生产实际的结论。这个问题应该使用因素各水平的均值之间协方差分析功能去解决。

② 使用系统默认值进行单因素方差分析只能得出是否有显著性差异的结论。本例数据量少，哪两组之间差别最大、哪种饲料使猪体重增加更快，几乎是可以看出来的。实际工作中，往往需要进行两两的组间均值比较，这就需要使用单因素 ANOVA 进行单因素方差分析时使用【两两比较】选项，从而获得更丰富的信息，使分析更深入。

③ 从主对话框可以看出，单因素方差分析允许分析一个因素变量对多个因变量的影响，主对话框中的【因变量列表】栏中可以移入多个因变量。

表 9-4 使用系统默认值的单因素方差分析结果

单因素方差分析					
猪体重增加量					
	平方和	df	均方	F	显著性
组间	20538.698	3	6846.233	157.467	.000
组内	652.159	15	43.477		
总数	21190.858	18			

9.2.2 单因素方差分析过程

单因素方差分析的选项分为 3 类，分别与 3 个功能按钮对应：对比功能可以指定一种要用 T 检验来检验的对比；两两比较功能可以指定一种多重比较检验；选项功能可以指定要输出的统计量，指定处理缺失值的方法。分别使用主对话框中的 3 个按钮打开相应的对话框，然后进行选择。

1. 进行对照比较的选项

在主对话框中，单击【对比】按钮，打开【单因素 ANOVA：对比】对话框，见图 9-4。在该对话框中可以把组间平方和划分成趋势成分或指定事先推测的对照比较。

(1) 趋势成分分析

考虑将组间偏差平方和分解为线性、二次、三次或更高次的趋势成分，操作如下：

① 选中【多项式】，该操作激活其右面的【度】(应为【阶次】)下拉菜单。

② 单因素 ANOVA 过程允许构造高达 5 次的均值多项式，多项式的阶数需要由读者自己根据研究的需要输入。单击【度】下拉菜单右面的向下箭头展开阶次菜单，可以选择的阶次有：线性、二次项、立方(三次项)、四次项、五次项。系统将在输出中给出指定阶次和低于指定阶次的各阶的平方和分解结果和各阶次的自由度、 $F$  值和  $F$  检验的概率值。



图 9-4 【单因素 ANOVA：对比】对话框

(2) 对照比较

对照比较在【1 的对比 1】栏中选择比较所需要的参数。可以选择多组比较参数。

① 系数指定规则。系数指定的顺序很重要，它应该与因素变量分组值的升序相对应。列表中第一个系数与因素变量最低一组的值相对应，而最后一个系数与因素变量最高组的值相对应。例如，如果因素变量有 6 个水平，系数列为-1、0、0、0、0.5、0.5，分别对应第一组到第六组。常用的是系数之和应该为 0，也可以设置系数之和不为 0，但会在输出中显示警告信息。在表 9-4 方差分析具有显著性意义的基础上，为检验假设  $H_0$ ：第一组均值与第四组均值间无显著差异(差异无统计意义)，可用系数 1, 0, 0, -1 来加以设定。

② 指定各组均值的系数具体的操作步骤为：在【系数】框中输入一个系数，单击【添加】按钮，输入的系数进入下面的方框中。重复上述操作，依次输入各组均值的系数，在方框中形成一系列数值。因素变量有几个水平(分为几组)，就输入几个系数，多出的无意义。不参与比较的分组系数应该为 0。如果多项式中只包括第一组与第四组的均值的系数，必须把第二个、第三个系数输入为“0”。如果只包括第一组与第二组的均值，则只需要输入前两个系数，第三个、第四个系数可以不输入。

可以同时进行多组均值组合比较。一组系数输入结束，单击【下…】按钮，系数框被清空，准备接受下一组系数数据。最多可以输入 10 组系数。

如果认为输入的几组系数中有错误，可以分别单击【上一张】按钮或【下…】按钮前后翻，找到出错的一组数据。单击出错的系数，该系数显示在编辑框中，可以在此进行修改或删除。修改后单击【更改】按钮，在系数显示框中出现正确的系数值。当在系数显示框中选中一个系数时，同时激活【删除】按钮，单击【删除】按钮删除所选中的系数。

2. 各组均值的多重成对比较选项

在主对话框中，单击【两两比较】按钮，展开【单因素 ANOVA：两两比较】对话框，见图 9-5。该对话框提供近 20 种组均值成对比较的检验，有些检验均值差异是否显著，给出差异一致性子集；有些成对进行组均值比较；有些进行这两种检验。成对比较产生的检验表多达 10 种。

非空组的组均值按升序排序，并用星号标明均值具有显著性差异的组对。另外，如果设计要求进行一致性子集检验，则计算一致性子集并将结果显示在一致性子集表中。

当各组观测数目不同时，除 R-E-G-WQ 和 R-E-G-WF 外，在计算一致性子集时，使用各组观测数目的调和均值作为各组样本含量。而这两个成对比较的检验，都是用各组本身的样本含量。

(1) 选择多重成对比较的方法。

① 各组方差齐性时在【假定方差齐性】栏中选择均值比较的方法，共 14 种方法、16 种选择。这些选项可以同时选择若干个，以便对各种均值比较方法的结果进行比较。

- **【LSD】**。用 T 检验完成各组均值间的成对比较，对多重比较误差率不进行调整。
- **【Bonferroni】**。计算 Student 统计量，完成各组间均值的成对比较。它通过设置每个检验的误差率来控制整个误差率。
- **【Sidak】**。计算  $t$  统计量进行成对比较，调整两两比较的显著性水平。限制比 Bonferroni 检验更严格。
- **【Scheffe】**。对所有可能的组合进行同步进入的配对比较。可以用于检验分组均数所有可能的线性组合。
- **【R-E-G-WF】**。用基于 F 检验的逐步缩小的成对比较的检验，显示一致性子集表。
- **【R-E-G-WQ】**。使用基于学生化值域逐步缩小的多元统计过程，进行子集一致性检验。
- **【S-N-K】**。使用学生化值域统计量，进行子集一致性检验。检验按均值递减排序，差异最大的先检验。
- **【Tukey】**。用 Student-Range 统计量进行所有组间均值进行配对比较，用所有配对比较的累计误差率作为试验误差率，还进行子集一致性检验。
- **【Tukey's-b】**。用学生化极差统计量进行组间均值的配对比较，其精确值为前两种检验相应值的平均值。
- **【Duncan】**。指定一系列的范围值，逐步进行计算比较得出结论，显示一致性子集检验结果。
- **【Hochberg's GT2】**。是基于学生化最大模数的检验。与 Tukey 类似，进行组均值成对比较和检测一致性子集。除非单元格含量非常不平衡，该检验甚至适用于方差不齐的情况。
- **【Gabriel】**。该方法根据学生化最大模数进行均值成对比较和子集一致性检验。当单元格含量不等时该方法比 Hochberg's GT2 更有效，在单元格含量较大时，这种方法较自由。
- **【Waller-Duncan】**。用  $t$  统计量进行子集一致性检验。使用贝叶斯逼近。
- **【Dunnett】**。使用 T 检验进行各组均值与对照组均值的比较。指定此选项，进行各组与对照组的均值比较，默认的对照组是最后一组。选择该项将激活下面的【控制类别】下拉列表，可以重新选择对照(控制)组为第一组。在被激活的【检验】栏中选择是进行双侧(尾)T 检验、各组均值是否都比对照组均值大的单尾 T 检验(【>控制】)，还是各组均值是否都比对照组均值小的单尾 T 检验(【<控制】)。



图 9-5 【单因素 ANOVA：两两比较】对话框

② 各组方差未检验是否具有齐性时,在【未假定方差齐性】(应为【假定方差不齐】)栏中选择检验各均数间是否有差异的方法,有 4 种可供选择:

- 【Tamhane's T2】。用 T 检验进行各组均值配对比较。
- 【Dunnett's T3】。用学生化最大模数检验进行各组均值间的配对比较。
- 【Games-Howell】。进行各组均值配对比较检验。该方法较灵活。
- 【Dunnett's C】。用学生化值域检验进行组均值配对比较。

③ 为便于选择,下面按功能列出。

- 进行均值两两比较的选项有【LSD】、【Sidak】、【Bonferroni】、【Games-Howell】、【Tamhane's T2】、【Dunnett's T3】、【Dunnett's C】、【Dunnett】(双尾、>控制、<控制)。
- 子集一致性检验的选项有【SNK】、【Tukey's-b】、【Duncan】、【R-E-G-WQ】、【R-E-G-WF】、【Waller-Duncan】。
- 进行均值两两比较和子集一致性检验两种检验的选项有【Hochberg's GT2】、【Tukey】、【Scheffe】、【Gabriel】。

(2) 【显著性水平】选项设定各种检验的显著性概率临界值,默认值为 0.05,可由读者重新设定。

### 3. 输出统计量的选择

在主对话框中,单击【选项】按钮,展开【单因素 ANOVA: 选项】对话框,见图 9-6。系统会按选择产生要求的统计量,并按要求的方式显示这些统计量。在该对话框中还可以选择对缺失值的处理要求。各组选项的含义如下:

① 【统计量】栏。

- 【描述性】。要求输出描述统计量:观测数目、均值、标准差、标准误、最小值、最大值、各组中每个因变量的 95%置信区间。
- 【固定和随机效果】。输出固定效应模型的标准差、标准误和 95%置信区间,以及随机效应模型的标准误、95%置信区间和方差成分间估测值。
- 【方差同质性检验】。要求进行方差齐性检验,并输出检验结果。用 Levene 检验计算每个观测与其组均值之差,然后对这些差值进行一维方差分析。
- 【Brown-Forsythe】。该统计量检验各组均数相等的,当不能确定方差齐性假设时,该统计量优于  $F$  统计量。
- 【Welch】。该统计量检验各组均数相等,当不能确定方差齐性假设是否成立时,该统计量优于  $F$  统计量。

② 【均值图】。要求作均数分布图,根据因素变量值所确定的各组均数描绘出因变量的均值分布情况。

③ 【缺失值】栏。选择缺失值处理方法。

- 【按分析顺序排除个案】。只有被选择参与分析的变量含缺失值的观测从分析中剔除。此为系统默认的处理方法。
- 【按列表排除个案】。对所有含有缺失值的观测从分析中剔除。

以上 3 组选项选择完成后,单击【继续】按钮,确认所作的选择并返回主对话框;单击【取消】按钮取消本次所做的所有选择,返回主对话框;单击【帮助】按钮,显示有关的帮助信息。



图 9-6 【单因素 ANOVA: 选项】对话框

9.2.3 单因素方差分析实例

【例 2】 分析不同饲料对猪体重的影响，见数据文件 data09-01。

(1) 按【分析→比较均值→单因素ANOVA】顺序打开主对话框。

(2) 指定因变量 *weight* (体重)、因素变量 *fodder* (饲料)。

(3) 指定选项。

① 单击【对比】按钮，打开相应的对话框，在【1 的对比 1】栏中指定了两组系数：

- 1、0、0、-1。检验 A、D 饲料对猪体重增加的效应及其之间是否有显著性差异。
- 0.5、-0.5、0.5、-0.5。检验 A、C 饲料之和效应是否与 B、D 之和效应有显著差异。

② 单击【两两比较】按钮，打开【单因素 ANOVA：两两比较】对话框，见图 9-5，选择均值配对比较的方法：在【假定方差齐性】栏中，选择【LSD】、【Duncan】两种方法；在【未假定方差齐性】栏中，选择【Tamhane's T2】方法；在【显著性水平】框中，输入“0.05”。单击【继续】按钮返回主对话框。

③ 单击【选项】按钮，打开【单因素 ANOVA：选项】对话框，见图 9-6，选择输出统计量：选中【描述性】，要求输出描述统计量；选中【方差同质性检验】，作方差齐性检验；选中【均值图】，要求作均值分布图；选中【按分析顺序剔除个案】，剔除参与分析的变量中有缺失值的观测。

(4) 输出结果见表 9-5～表 9-11、图 9-7。

表 9-5 描述统计量

猪体重增加量								
	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
1 A	5	133.3600	6.80794	3.04460	124.9068	141.8132	125.30	143.10
2 B	5	152.0400	6.95723	3.11137	143.4015	160.6785	143.80	162.70
3 C	5	189.7200	6.35035	2.83996	181.8350	197.6050	182.80	198.60
4 D	4	220.7750	6.10594	3.05297	211.0591	230.4909	212.30	225.80
总数	19	171.5105	34.31137	7.87157	154.9730	188.0481	125.30	225.80

表 9-6 方差齐性检验结果

猪体重增加量			
Levene 统计量	df1	df2	显著性
.024	3	15	.995

表 9-7 单因素方差分析结果表

猪体重增加量					
	平方和	df	均方	F	显著性
组间	20538.698	3	6846.233	157.467	.000
组内	652.159	15	43.477		
总数	21190.858	18			

表 9-8 对比系数

对比	饲料			
	1 A	2 B	3 C	4 D
1	1	0	0	-1
2	.5	-.5	.5	-.5

(5) 结果说明。

表 9-5 所示为描述统计量结果，给出了 4 种饲料分组的样本含量 N、因变量猪体重的平均数、标准差、标准误、95%的置信区间上下限以及最小和最大值。

表 9-6 所示为方差齐性检验结果。从  $\text{sig.} = 0.995$  得出  $p > 0.05$ ，说明各组的方差在  $\alpha = 0.05$  水平上没有显著性差异，即方差具有齐性。

表 9-9 对比检验

		对比	对比值	标准误	t	df	显著性 (双 侧)
猪体重增加量	假设方差相等	1	-87.4150	4.42321	-19.763	15	.000
		2	-24.8675	3.03956	-8.181	15	.000
	不假设等方差	1	-87.4150	4.31164	-20.274	6.852	.000
		2	-24.8675	3.01398	-8.251	14.649	.000

表 9-7 所示是方差分析结果。给出了组间、组内的偏差平方和、均方、 $F$  值和显著性概率  $p$  值。 $p < 0.05$ , 各组间均值在  $\alpha = 0.05$  水平上有显著性差异。

表 9-8 所示为对比系数表, 列出两组均值对比的系数, 用以检查对比目的是否表达正确。

表 9-9 所示为均值对比结果, 表中内容解释如下。

第一栏: 按方差齐性和非齐性划分。表 9-5 已得出方差具有齐性的结论, 所以选择假设方差相等(方差齐性)一行的数据得出结论。

第二栏: 结合表 9-8 和表 9-7 得出该栏数据。第一个对比检验的是 A 组和 D 组均值是否有显著性差异, 两组均值之差为 -87.415 为 A-D 的值; 第二个对比值为 -49.735, 是  $0.5A - 0.5B - 0.5C + 0.5D$  的计算结果, 其中大写字母代表各组因变量均值。

第三栏: 标准误。

第四栏: 计算的  $t$  值, 是第二栏与第三栏之比。

表 9-10 均值多重比较的结果

因变量: 猪体重增加量

		(I) 饲料	(J) 饲料	均值差 (I-J)	标准误	显著性	95% 置信区间	
							下限	上限
LSD	1 A	2 B		-18.68000 <sup>*</sup>	4.17024	.000	-27.5687	-9.7913
		3 C		-56.36000 <sup>*</sup>	4.17024	.000	-65.2487	-47.4713
		4 D		-87.41500 <sup>*</sup>	4.42321	.000	-96.8428	-77.9872
	2 B	1 A		18.68000 <sup>*</sup>	4.17024	.000	9.7913	27.5687
		3 C		-37.68000 <sup>*</sup>	4.17024	.000	-46.5687	-28.7913
		4 D		-68.73500 <sup>*</sup>	4.42321	.000	-78.1628	-59.3072
	3 C	1 A		56.36000 <sup>*</sup>	4.17024	.000	47.4713	65.2487
		2 B		37.68000 <sup>*</sup>	4.17024	.000	28.7913	46.5687
		4 D		-31.05500 <sup>*</sup>	4.42321	.000	-40.4828	-21.6272
	4 D	1 A		87.41500 <sup>*</sup>	4.42321	.000	77.9872	96.8428
		2 B		68.73500 <sup>*</sup>	4.42321	.000	59.3072	78.1628
		3 C		31.05500 <sup>*</sup>	4.42321	.000	21.6272	40.4828
Tamhane	1 A	2 B		-18.68000 <sup>*</sup>	4.35318	.016	-33.7633	-3.5967
		3 C		-56.36000 <sup>*</sup>	4.16353	.000	-70.8053	-41.9147
		4 D		-87.41500 <sup>*</sup>	4.31164	.000	-103.1431	-71.6869
	2 B	1 A		18.68000 <sup>*</sup>	4.35318	.016	3.5967	33.7633
		3 C		-37.68000 <sup>*</sup>	4.21260	.000	-52.3109	-23.0491
		4 D		-68.73500 <sup>*</sup>	4.35904	.000	-84.6022	-52.8678
	3 C	1 A		56.36000 <sup>*</sup>	4.16353	.000	41.9147	70.8053
		2 B		37.68000 <sup>*</sup>	4.21260	.000	23.0491	52.3109
		4 D		-31.05500 <sup>*</sup>	4.16966	.001	-46.4051	-15.7049
	4 D	1 A		87.41500 <sup>*</sup>	4.31164	.000	71.6869	103.1431
		2 B		68.73500 <sup>*</sup>	4.35904	.000	52.8678	84.6022
		3 C		31.05500 <sup>*</sup>	4.16966	.001	15.7049	46.4051
Dunnett t (双侧) <sup>b</sup>	1 A	4 D		-87.41500 <sup>*</sup>	4.42321	.000	-98.8910	-75.9390
	2 B	4 D		-68.73500 <sup>*</sup>	4.42321	.000	-80.2110	-57.2590
	3 C	4 D		-31.05500 <sup>*</sup>	4.42321	.000	-42.5310	-19.5790

\*. 均值差的显著性水平为 0.05。

b. Dunnett t 检验将一个组视为一个控制组, 并将其与所有其他组进行比较。



表 9-11 DUNCAN 法一致性子集检验结果

猪体重增加量						
饲料		N	alpha = 0.05 的子集			
			1	2	3	4
Duncan <sup>a,b</sup>	1 A	5	133.3600			
	2 B	5		152.0400		
	3 C	5			189.7200	
	4 D	4				220.7750
显著性			1.000	1.000	1.000	1.000

将显示同类子集中的组均值。  
a. 将使用调和均值样本大小 = 4.706。  
b. 组大小不相等。将使用组大小的调和均值。将不保证 I 类错误级别。

第五栏：df 自由度。

第六栏：双侧检验的显著性概率。从概率值可以看出：对比 1,  $p < 0.05$ ；对比 2,  $p < 0.05$ 。因此饲料对猪体重增加的效应，A、D 效应均值之间在  $\alpha = 0.05$  水平上有显著性差异。而 A、C 之和效应与 B、D 之和效应之间有显著性差异。从对比值栏内值的符号和描述统计表中均值栏内的数据不难得出各对比组均值之差。

表 9-10 所示是 LSD 法和 Tamhane’s T2 法进行均值多重比较的结果。从选择比较方法处知 LSD 属于假定方差齐性栏的选项，从表 9-5 得知方差具有齐性，因此只需从 LSD 法结果作结论。比较结果说明，A 与 B、A 与 C、A 与 D、B 与 C、B 与 D、C 与 D 各组均值间均有显著性差异。表中 “\*” 标示的组均值在  $\alpha = 0.05$  水平上有显著性差异。

表 9-11 所示为一致性子集检验结果。第一栏列出 A、B、C、D 各组。第二栏列出各组观测数。由于各组样本含量不等，计算均数用的是调和平均数的样本量，为 4.706。各组猪体重增加量的均值单独为一个子集，说明没有两组均值相等的情况。与多重比较结果一致。

图 9-7 所示是以因素变量“饲料”为横轴，以独立变量“猪体重增加量的均值”为纵轴绘制的均值散点图。可直观地看出各组均值的分布。

应该特别说明的是，选取哪些选项是根据研究需要进行的。本例中希望比较各种饲料对猪体重增加的效应，因此选择多重比较的选项。两个均值组合对比在此例中可能无实际意义，只是为了说明选项的使用方法才选择了该项。

【例 3】 方差分析不同细菌对三叶草含氮量的影响。

本例是 Erdman(1946)的一个试验，同种三叶草被接种上不同的菌种测量三叶草植物中含氮量。每组数据中的前面一个是菌种代码，变量名是 strain。SPSS 分析过程要求因素变量必须为数值型变量。后面一个是含氮量，变量名是 nitrogen。数据文件为 data09-02。

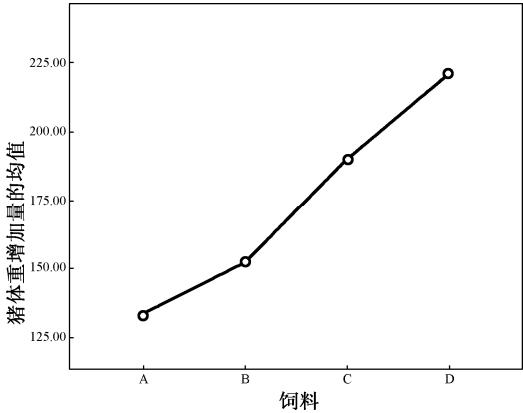


图 9-7 均值散点图

(1) 操作简述。

- ① 调用单因素 ANOVA 过程：【分析→比较均值→单因素 ANOVA】。
- ② 在主对话框中指定因变量 nitrogen、因素变量 strain。
- ③ 在主对话框中单击【选项】按钮，选择输出描述统计量和方差同质性(齐性)检验结果。

- ④ 使用系统默认方法处理缺失值。
- ⑤ 在主对话框中单击【两两比较】按钮，要求进行均值多重比较，采用【LSD】、【TUKEY】、【Tamhane's T2】方法；显著性概率临界值设定为“0.05”。
- (2) 输出结果见表 9-12～表 9-16。

表 9-12 描述统计量

三叶草含氮量									
	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值	
					下限	上限			
1	5	28.820	5.8002	2.5939	21.618	36.022	19.4	33.0	
4	5	14.640	4.1162	1.8408	9.529	19.751	9.1	19.4	
5	5	23.980	3.7772	1.6892	19.290	28.670	17.7	27.9	
7	5	19.920	1.1300	.5054	18.517	21.323	18.6	21.0	
13	5	13.260	1.4276	.6384	11.487	15.033	11.6	14.4	
30	5	18.700	1.6016	.7162	16.711	20.689	16.9	20.8	
总数	30	19.887	6.2422	1.1397	17.556	22.218	9.1	33.0	

表 9-13 方差齐性检验结果

三叶草含氮量			
Levene 统计量	df1	df2	显著性
3.145	5	24	.025

表 9-14 单因素方差分析

三叶草含氮量					
	平方和	df	均方	F	显著性
组间	847.047	5	169.409	14.371	.000
组内	282.928	24	11.789		
总数	1129.975	29			

表 9-15 多重比较结果(一张表的三部分)

因变量: 三叶草含氮量									
	(I) 菌株编号	(J) 菌株编号	均值差 (I-J)	标准误	显著性	95% 置信区间			
						下限	上限		
Tukey HSD	1	4	14.1800 <sup>*</sup>	2.1715	.000	7.466	20.894		
		5	4.8400	2.1715	.262	-1.874	11.554		
		7	8.9000 <sup>*</sup>	2.1715	.005	2.186	15.614		
		13	15.5600 <sup>*</sup>	2.1715	.000	8.846	22.274		
		30	10.1200 <sup>*</sup>	2.1715	.001	3.406	16.834		
	4	1	-14.1800 <sup>*</sup>	2.1715	.000	-20.894	-7.466		
		5	-9.3400 <sup>*</sup>	2.1715	.003	-16.054	-2.626		
		7	-5.2800	2.1715	.185	-11.994	1.434		
		13	1.3800	2.1715	.987	-5.334	8.094		
		30	-4.0600	2.1715	.443	-10.774	2.654		
	5	1	-4.8400	2.1715	.262	-11.554	1.874		
		4	9.3400 <sup>*</sup>	2.1715	.003	2.626	16.054		
		7	4.0600	2.1715	.443	-2.654	10.774		
		13	10.7200 <sup>*</sup>	2.1715	.001	4.006	17.434		
		30	5.2800	2.1715	.185	-1.434	11.994		
	7	1	-8.9000 <sup>*</sup>	2.1715	.005	-15.614	-2.186		
		4	5.2800	2.1715	.185	-1.434	11.994		
		5	-4.0600	2.1715	.443	-10.774	2.654		
		13	6.6600	2.1715	.053	-0.054	13.374		
		30	1.2200	2.1715	.993	-5.494	7.934		
	13	1	-15.5600 <sup>*</sup>	2.1715	.000	-22.274	-8.846		
		4	-1.3800	2.1715	.987	-8.094	5.334		
		5	-10.7200 <sup>*</sup>	2.1715	.001	-17.434	-4.006		
		7	-6.6600	2.1715	.053	-13.374	.054		
		30	-5.4400	2.1715	.162	-12.154	1.274		
	30	1	-10.1200 <sup>*</sup>	2.1715	.001	-16.834	-3.406		
		4	4.0600	2.1715	.443	-2.654	10.774		
		5	-5.2800	2.1715	.185	-11.994	1.434		
		7	-1.2200	2.1715	.993	-7.934	5.494		
		13	5.4400	2.1715	.162	-1.274	12.154		

(续表)

LSD	1	4	14.1800 <sup>*</sup>	2.1715	.000	9.698	18.662
		5	4.8400 <sup>*</sup>	2.1715	.035	.358	9.322
		7	8.9000 <sup>*</sup>	2.1715	.000	4.418	13.382
		13	15.5600 <sup>*</sup>	2.1715	.000	11.078	20.042
		30	10.1200 <sup>*</sup>	2.1715	.000	5.638	14.602
	4	1	-14.1800 <sup>*</sup>	2.1715	.000	-18.662	-9.698
		5	-9.3400 <sup>*</sup>	2.1715	.000	-13.822	-4.858
		7	-5.2800 <sup>*</sup>	2.1715	.023	-9.762	-.798
		13	1.3800	2.1715	.531	-3.102	5.862
		30	-4.0600	2.1715	.074	-8.542	.422
	5	1	-4.8400 <sup>*</sup>	2.1715	.035	-9.322	-.358
		4	9.3400 <sup>*</sup>	2.1715	.000	4.858	13.822
		7	4.0600	2.1715	.074	-.422	8.542
		13	10.7200 <sup>*</sup>	2.1715	.000	6.238	15.202
		30	5.2800 <sup>*</sup>	2.1715	.023	.798	9.762
	7	1	-8.9000 <sup>*</sup>	2.1715	.000	-13.382	-4.418
		4	5.2800 <sup>*</sup>	2.1715	.023	.798	9.762
		5	-4.0600	2.1715	.074	-8.542	.422
		13	6.6600 <sup>*</sup>	2.1715	.005	2.178	11.142
		30	1.2200	2.1715	.579	-3.262	5.702
	13	1	-15.5600 <sup>*</sup>	2.1715	.000	-20.042	-11.078
		4	-1.3800	2.1715	.531	-5.862	3.102
		5	-10.7200 <sup>*</sup>	2.1715	.000	-15.202	-6.238
		7	-6.6600 <sup>*</sup>	2.1715	.005	-11.142	-2.178
		30	-5.4400 <sup>*</sup>	2.1715	.019	-9.922	-.958
	30	1	-10.1200 <sup>*</sup>	2.1715	.000	-14.602	-5.638
		4	4.0600	2.1715	.074	-.422	8.542
		5	-5.2800 <sup>*</sup>	2.1715	.023	-9.762	-.798
		7	-1.2200	2.1715	.579	-5.702	3.262
		13	5.4400 <sup>*</sup>	2.1715	.019	.958	9.922

Tamhane	1	4	14.1800 <sup>*</sup>	3.1807	.040	.569	27.791
		5	4.8400	3.0954	.930	-8.690	18.370
		7	8.9000	2.6427	.317	-6.522	24.322
		13	15.5600 <sup>*</sup>	2.6713	.044	.471	30.649
		30	10.1200	2.6910	.206	-4.768	25.008
	4	1	-14.1800 <sup>*</sup>	3.1807	.040	-27.791	-.569
		5	-9.3400	2.4984	.083	-19.625	.945
		7	-5.2800	1.9089	.485	-15.853	5.293
		13	1.3800	1.9484	1.000	-8.860	11.620
		30	-4.0600	1.9752	.769	-14.125	6.005
	5	1	-4.8400	3.0954	.930	-18.370	8.690
		4	9.3400	2.4984	.083	-.945	19.625
		7	4.0600	1.7632	.678	-5.535	13.655
		13	10.7200 <sup>*</sup>	1.8058	.026	1.444	19.996
		30	5.2800	1.8348	.384	-3.839	14.399
	7	1	-8.9000	2.6427	.317	-24.322	6.522
		4	5.2800	1.9089	.485	-5.293	15.853
		5	-4.0600	1.7632	.678	-13.655	5.535
		13	6.6600 <sup>*</sup>	.8142	.001	3.250	10.070
		30	1.2200	.8766	.968	-2.536	4.976
	13	1	-15.5600 <sup>*</sup>	2.6713	.044	-30.649	-.471
		4	-1.3800	1.9484	1.000	-11.620	8.860
		5	-10.7200 <sup>*</sup>	1.8058	.026	-19.996	-1.444
		7	-6.6600 <sup>*</sup>	.8142	.001	-10.070	-3.250
		30	-5.4400 <sup>*</sup>	.9595	.007	-9.398	-1.482
	30	1	-10.1200	2.6910	.206	-25.008	4.768
		4	4.0600	1.9752	.769	-6.005	14.125
		5	-5.2800	1.8348	.384	-14.399	3.839
		7	-1.2200	.8766	.968	-4.976	2.536
		13	5.4400 <sup>*</sup>	.9595	.007	1.482	9.398

\*. The mean difference is significant at the 0.05 level.

表 9-16 一致性子集检验结果

三叶草含氮量					
	菌株编号	N	alpha = 0.05 的子集		
			1	2	3
Tukey HSD <sup>a</sup>	13	5	13.260		
	4	5	14.640		
	30	5	18.700	18.700	
	7	5	19.920	19.920	
	5	5		23.980	23.980
	1	5			28.820
显著性			.053	.185	.262

将显示同类子集中的组均值。

a. 将使用调和均值样本大小 = 5,000。

本例输出与例 2 的输出格式一致，读者可以自己从中得出结论。需要注意的是，表 9-13 的方差齐性检验得出方差不具有齐性的结论，在进行多重比较时应选择 Tamhane 方法作结论。从 Tamhane 方法的结果看，1 与 4、1 与 13、5 与 13、7 与 13、13 与 30 菌种之间的含氮量均值差异是有统计意义的。表 9-16 所示是指定 Tukey HSB 方法而产生的子集一致性检验结果。因为该方法要求方差具有齐性，故在本例中不适用，在实际应用中，不必列出本方法的结果。

## 9.3 单因变量多因素方差分析

### 9.3.1 单因变量多因素方差分析概述

#### 1. 概述

单因变量多因素方差分析是对一个独立变量是否受多个因素或变量影响而进行的方差分析。SPSS 调用【单因变量多因素方差分析】(UNIANOVA)过程, 检验不同水平组合之间因变量均数由于受不同因素影响是否有差异的问题。在这个过程中可以分析每一个因素的作用, 也可以分析因素之间的交互作用。可以进行协方差分析, 以及各因素变量与协变量之间的交互作用。该过程要求因变量从多元正态总体随机采样得来, 且总体中各单元的方差相同, 也可以通过方差齐性检验选择均值比较结果。

因变量和协变量必须是数值型变量, 协变量与因变量彼此不独立。因素变量是分类变量, 可以是数值型, 也可以是长度不超过 8 的字符型变量。固定因素变量(Fixed Factor)反应处理的因素。随机因素是随机设置的因素, 是在确定模型时需要考虑会对实验有影响的因素, 对试验结果影响的大小可以通过方差成分分析确定。

#### 2. 关于模型

单因变量一般线性模型功能很强, 可以建立包括各种主效应、交互效应的模型。必须认真分析因素变量的具体情况, 来确定自己的模型, 否则会产生不可解释的输出结果。

### 9.3.2 单因变量多因素方差分析过程

单因变量多因素方差分析的功能模块调用步骤见 9-2。即按【分析→一般线性模型→单变量】的顺序打开【单变量】主对话框, 见图 9-8。

用与 9.2 节中叙述的相同方法确定因变量, 将因变量移到【因变量】框中。定义固定因素变量, 并将其移到【固定因子】框中。将随机因素变量移到【随机因子】框中。

**注意:** 由于内存容量的限制, 选择的因素水平组合数(单元数)应该尽量少。因素数量和对选定因素定义的取值数量决定了组合数。

如果需要去除协变量的影响, 则将协变量移到【协变量】框中。

【WLS 权重】框允许指定一个权重变量, 用于加权的最小平方分析。权重变量给观测不同的权重, 也可以给不同测量精度以不同的补偿。如果需要考虑权重变量的影响, 将权重变量移到【WLS 权重】框中。

可通过功能按钮展开相应对话框选择【模型】、【对比】和选择输出统计量。

#### 1. 选择分析模型

在主对话框中, 单击【模型】按钮, 打开【单变量: 模型】对话框, 见图 9-9。

(1) 在【指定模型】栏中, 指定模型类型。

①【全因子】为系统默认的模式, 即全模型。全模型包括所有因素变量的主效应、所有协变量主效应、所有因素与因素的交互效应, 不包括协变量与其他因素的交互效应。

不打开此对话框, 即选择了全模型。



图 9-8 【单变量】主对话框

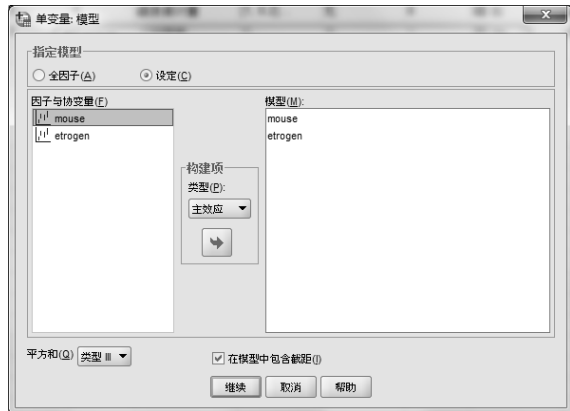


图 9-9 【单变量: 模型】对话框

② 【设定】。建立自定义的模型。此项的选择激活下面各操作框。

(2) 建立自定义模型。

选择了【设定】选项后，在【因子与协变量】框中自动列出可以作为因素变量的变量名，根据表中列出的变量名建立模型。

① 选择模型中的主效应。在【因子与协变量】框中选择一个因素变量名，单击【构建项】栏中下面的箭头，送入【模型】框中，一个变量名占一行，称为主效应项。欲在模型中包括几个主效应项，就进行几次以上的操作。也可以选择多个一次送入模型框中。

② 选择交互效应类型。单击【构建项】框，【类型】下面的选项内容框中单击右面的向下箭头可以展开下拉菜单。其中有如下几项：

- 【主效应】。选择该项只可以指定主效应。
- 【交互】。选中此项可以指定任意的交互效应。
- 【所有二阶】、【所有三阶】、【所有四阶】和【所有五阶】选项。指定所有二阶交互效应到所有五阶交互效应。在下拉菜单中单击某一项，选中的交互类型显示在矩形框中。

③ 建立模型中的交互项。以 3 个因素变量为例，方法如下：

- 要求模型中包括两个变量的二阶交互效应。相应的操作是在【因子与协变量】框内的变量表中选择一个变量，此为选择了交互项之一，再选择第二个变量，此为选择了交互项之二。单击【构建项】栏内参数框的箭头按钮，一个交互效应出现在【模型】框中。模型增加了一个交互效应项：两个变量名之间用“\*”连接。
- 要求模型中包括 3 个变量的所有二阶交互效应项时应该分别单击 3 个变量名。在【构建项】栏内参数框中选择【所有二阶】项，单击箭头按钮。在模型框中出现 3 个二阶交互效应项：两两变量名用“\*”连接的表达式共 3 个。
- 若要求模型中包括所有三阶效应，分 3 次单击 3 个变量，选择【构建项】栏内参数框中的【交互】或【所有三阶】项，再单击箭头按钮，均可以在【模型】框中出现三维交互效应项：3 个变量名间用“\*”连接。

(3) 选择分解平方和的方法。

在对话框的下部有【平方和】选项框，可以进行四项选择来确定平方和的分解方法：【类型 I】、【类型 II】、【类型 III】和【类型 IV】。其中，【类型 III】是系统默认的，也是常用的一种方法。

①【类型 I】。分层处理平方和的方法。仅对模型主效应之前的每项进行调整。一般适用于：平衡的 ANOVA 模型，在这个模型中一阶交互效应前指定主效应，二阶交互效应前指定一阶交互效应，依次类推；多项式回归模型中，任何低阶项都在较高阶项前面指定；完全嵌套模型，在模型中第一个被指定的效应嵌套在第二个被指定的效应中，第二个被指定的效应嵌套在第三个被指定的效应中，嵌套模型只能使用语句指定。

②【类型 II】。该方法计算一个效应的平方和时，对其他所有的效应进行调整。一般适用于平衡的 ANOVA 模型、仅有主效应的模型、任何回归模型、完全嵌套设计。

③【类型 III】。是系统默认的处理方法，对其他任何效应均进行调整。它的优势是把所估计的剩余常量也考虑到单元频数中。一般适用于：类型 I、类型 II 所列的模型和没有空单元格的平衡和不平衡模型。

④【类型 IV】。该方法是为有缺失单元格的情况设计的。使用此方法对任何效应 F 计算平方和。如果 F 不包含在其他效应里，类型 IV=类型 III=类型 II；如果 F 包含在其他效应里，类型 IV 只对 F 的较高水平效应参数作对比。一般适用于：类型 I、类型 II 所列模型和有空单元格的平衡和不平衡模型。

(4) 选中【在模型中包括截距】。系统默认截距包括在回归模型中。如果能假设数据通过原点，可以不包括截距，就不选择此项。

## 2. 选择对照方法

在主对话框中，单击【对比】按钮，打开【单变量：对比】对话框，见图 9-10。



图 9-10 【单变量：对比】对话框

(1) 在【因子】框中显示出所有在主对话框中选中的因素变量。因素变量名后的括号中是当前的对比方法。默认的是不进行对比，即显示“无”。

(2) 在【更改对比】栏中改变对照方法。对比检验一个因素的各水平间的差异。可以对模型中的每个因素指定一种对比方法，对比结果描述的是参数的线性组合。操作方法如下：

① 在【因子】框中选择想要改变对照方法的因子，激活【更改对比】栏中的各项。

② 单击【对比】参数框中的向下箭头，在展开的对照方法列表中选择对照方法，可供选择的对照方法包括：

- 【无】。不进行均数比较。
- 【偏差】。除被忽略的水平外，比较因素变量(或称预测变量)的每个水平的效应。可以选择最后一个水平或第一个水平作为忽略的水平。
- 【简单】。除了作为参考的水平外，对预测变量或因素变量的每一水平都与参考水平进行比较。选择最后一个水平或第一个水平作为参考水平。
- 【差值】。对预测变量(或因素)的每一水平的效应，除第一水平以外，都与其前面各水平的平均效应进行比较。与 Helmert 对照方法相反。
- 【Helmert】。对因素的效应，除最后一个以外，都与后续的各水平的平均效应相比较。
- 【重复】。对相邻的水平进行比较。对因素的效应，除第一水平以外，对每一水平都与其前面的水平进行比较。

- **【多项式】**。第一级自由度包括线性效应与预测变量或因素水平的交叉，第二级包括二次效应等。各水平彼此的间隔被假设是均匀的。

③ 单击**【更改】**按钮，选中的(或改变了的)对照方法显示在步骤①选中的因子变量后面的括号中。

④ 只有选择了**【偏差】**或**【简单】**方法时才需要选择**【参考水平】**(SPSS 汉化为**【参考类别】**)。共有两种参考水平可选择：**【最后一个】**水平和**【第一个】**水平。系统默认的参考水平是**【最后一个】**。

### 3. 选择分布图形

在主对话框中单击**【绘制】**按钮，打开**【单变量：轮廓图】**对话框，见图 9-11。在该对话框中，选择作边际均值图的参数。

边际均值图(Profile)用于比较边际均值。边际均值图是线图，图中每个点表明因变量在因素变量每个水平上的边际均值的估计值。如果指定了协变量，该均值则是经过协变量调整的均值。纵轴是因变量；横轴是一个因素变量。

作单因素方差分析时，边际均值图表明该因素各水平的因变量均值。

双因素方差分析时，指定一个因素作横轴变量，另一个因素变量的每个水平产生不同的线。如果是三因素方差分析，可以指定第三个因素变量，该因素每个水平产生一个边际均值图。双因素或多因素边际均值图中相互平行的线表明在因素间无交互效应，不平行的线表明因素间存在交互效应，见图 9-12 和图 9-13。具体操作如下：

(1) **【因子】**框中为主对话框中所选因素变量名。

(2) **【水平轴】**框。选择**【因子】**框中一个因素变量作水平轴变量，单击箭头按钮，将其送入相应的水平轴轴框中。

如果只想看该因素变量各水平的因变量均值分布，单击**【添加】**按钮，将所选因素变量移入下面的**【图】**框中；否则，不单击**【添加】**按钮，接着进行下一步。



图 9-11 **【单变量：轮廓图】**对话框

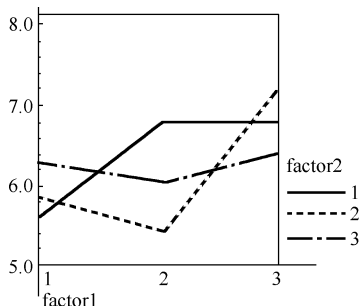


图 9-12 两因素变量有交互作用

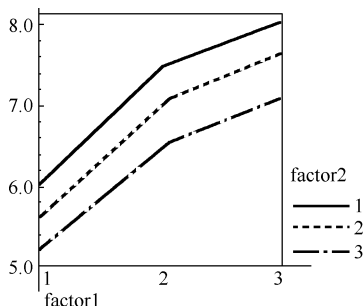


图 9-13 两因素变量无交互作用

(3) **【单图】**框。确定分线变量。如果想看两个因素变量组合的各单元格中因变量均值分布，或想看两个因变量间是否存在交互效应，选择**【因子】**框中另一个因素变量，单击箭头按钮，将变量名送入**【单图】**框中。单击**【添加】**按钮，将自动生成的图形表达式送入**【图】**栏

中。【单图】框中变量的每个水平在图中是一条线。图形表达式是用“\*”连接的两个因素变量名。

(4) 【多图】框。确定分图变量。如果在【因子】栏中还有因素变量，可以按上述方法，将其送入【多图】框中作为分图变量，单击【添加】按钮，将自动生成的图形表达式送入【图】栏中。图形表达式是用“\*”连接的 3 个因素变量名。分图变量的每个水平生成一张线图。

(5) 如果将图形表达式送到【图】框后发现错误，可以修改和删除。单击有错的图形表达式，该表达式所包括的变量显示在输入的位置。对选错的变量，将其送回源变量框中，再重新输入正确内容，然后单击【更改】按钮改变表达式，检查无误后，单击【继续】按钮确认，返回到主对话框。

#### 4. 选择多重比较分析

在主对话框中，单击【两两比较】按钮，打开【单变量：观测均值的两两比较】对话框。从【因子】框中选择变量，单击箭头按钮，使被选变量进入【两两比较】检验框，然后选择多重比较方法，方法的选择参见 9.2.2 节内容。

#### 5. 保存运算结果的选项

在主对话框中，单击【保存】按钮，打开【单变量：保存】对话框，见图 9-14。通过在对话框中的选择，系统使用默认变量名将所计算的预测值、残差值和诊断值作为新的变量保存在当前数据文件中。以便于在进一步的统计分析中使用这些值。在数据编辑窗中，用鼠标指向变量名，会显示对该新生成变量含义的解释。

(1) 【预测值】栏。系统对每个观测给出根据模型计算的预测值。

① 【未标准化】。给出非标准化预测值。

② 【加权】。如果在主对话框中选择了 WLS 加权变量，选中该项将保存加权的非标准化预测值。

③ 【标准误】。给出预测值标准误。

(2) 【诊断】栏。测量并标识对模型影响较大的观测或自变量。根据选择还可以给出：

① 【Cook 距离】。

② 【杠杆值】。给出非中心化杠杆值。

(3) 【残差】栏。

① 【未标准化】。给出未标准化残差值，即观测值与预测值之差。

② 【加权】。如果在主对话框中选择了 WLS 加权变量，选中该项将保存加权的未标准化残差。

③ 【标准化】。给出标准化残差，又称 Pearson 残差。

④ 【学生化】。给出学生化残差。

⑤ 【剔除】(SPSS 汉化为【删除】)。给出剔除残差，即因变量值与校正预测值之差。

以上选项给出的有关回归的统计量含义请参考第 11 章的有关内容。

(4) 【系数统计】栏。选中【系数统计】栏中的【创建系数统计】项，可将模型参数估计的方差-协方差矩阵保存到一个新文件中。对因变量产生三行数据：一行是参数估计值，一行

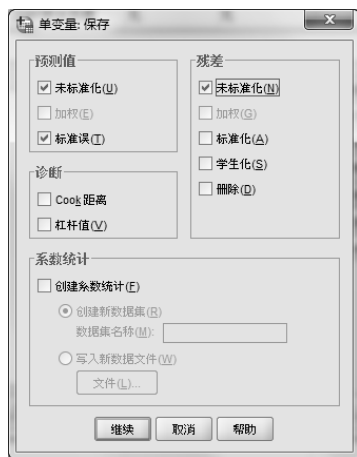


图 9-14 【单变量：保存】对话框



是与参数估计值相对应的显著性检验的  $t$  统计量，还有一行是残差自由度。数据可以有两种处理方式：

① **【创建新数据集】**。选择该项，需要给出新数据集名称。

② **【写入新数据文件】**。所生成的新数据文件可以作为另外分析的输入数据文件。单击 **【文件】** 按钮，打开相应的保存对话框，指定文件的保存位置和文件名。

## 6. 选择输出项

在主对话框中，单击 **【选项】** 按钮，打开 **【单变量：选项】** 对话框，见图 9-15。

(1) **【估计的边际均值】** 栏。

① **【因子与因子交互】** 框中列出了在模型对话框中所指定的效应项。在该框中选定因素变量的各种效应项，单击移动箭头，将其复制到显示均值框中。选择主效应，则产生估计的边际均值表。选择二维交互效应产生的估计边际均值表实际上是典型的单元格均值表。选择三维交互效应也显示单元格均值表。选择 **【OVERALL】** 项产生边际均值的均值。详见第 9.3.4 节中的例 5。

② 在 **【显示均值】** 框中有主效应时激活此框下面的 **【比较主效应】** 复选项，对主效应的边际均值进行组间的配对比较。

③ **【置信区间调节】** 框的下拉列表中列出了进行了多重组间比较时置信区间和显著性水平调整方法的选项，共有 3 个选项：

- **【LSD (none)】**。不进行调整。
- **【Bonferroni】**。邦弗伦尼方法，是基于 Student  $t$  统计量的方法。适用于要进行比较的均值，对数比较少少的情况。
- **【Sidak】**。计算  $t$  统计量进行多重配对比较，调整多重比较的显著性水平。其限制比 Bonferroni 检验更严格。

(2) **【输出】** 栏。指定要输出的统计量。

① **【描述统计】**。输出的描述统计量，有观测均值、标准差和各单元格中的观测数。

② **【功效估计】**。输出效应量估计，给出  $\eta^2$  (eta square)。它反映了每个效应与每个参数估计值可以归于因素的总变异的大小。

③ **【检验效能】**。给出各种检验假设的功效。计算功效的显著性水平，系统默认的临界值是 0.05。

④ **【参数估计】**。给出各因素变量的模型参数估计、标准误、T 检验的  $t$  值、显著性概率和 95% 的置信区间。

⑤ **【对比系数矩阵】**。显示变换系数矩阵或 L 矩阵。

⑥ **【方差齐性检验】**。

⑦ **【分布-水平图】**。绘制观测均值-标准差图、观测均值-方差图。

⑧ **【残差图】**。绘制残差图，给出观测值、预测值散点图和观测数目对标准化残差的散点图，加上正态和标准化残差的正态概率图。

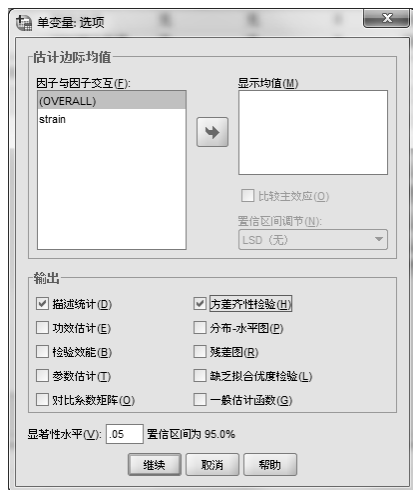


图 9-15 **【单变量：选项】** 对话框

⑨【缺乏拟合优度检验】(【失拟】(lack of fit))。检查独立变量和非独立变量间的关系是否被充分描述。

⑩【一般估计函数】。可以根据一般估计函数自定义假设检验。对比系数矩阵的行与一般估计函数是线性组合的。

(3) 在【显著性水平】框中改变多重比较的显著性水平，并给出置信区间。

9.3.3 随机区组设计的方差分析实例

【例 4】 4 个种系未成年雌性大白鼠各 3 只，每只按一种剂量注射雌激素，一段时间后，解剖称子宫重量。数据见表 9-17，数据录入格式见表 9-16，数据文件为 data09-03。

1) 操作方法与步骤

(1) 定义 3 个变量，建立数据文件：2 个分类变量，1 个尺度(连续)变量。

- ① 大白鼠种系变量 mouse，取值 1~4，是种系 A~D 的代码。
- ② 雌激素剂量变量 etrogen，取值 1~3，是剂量 0.2、0.4、0.8 三种剂量的代码。
- ③ 子宫重量变量 wuteri，连续变量，是本例的研究对象。

输入数据时应该注意观测是如何构成的。正确的构成方式应该如图 9-16 所示。

表 9-17 不同种系、剂量的子宫重量

种系	剂 量		
	0.2 (1)	0.4 (2)	0.8 (3)
A (1)	106	116	145
B (2)	42	68	115
C (3)	70	111	133
D (4)	42	63	87

	mouse	etrogen	wuteri
1	1	1	106
2	1	2	116
3	1	3	145
4	2	1	42
5	2	2	68
6	2	3	115
7	3	1	70
8	3	2	111
9	3	3	133
10	4	1	42
11	4	2	63
12	4	3	87

图 9-16 方差分析的数据安排

(2) 按【分析→一般线性模型→单变量】顺序单击菜单项，打开单变量主对话框，见图 9-8。

(3) 定义因变量和因素变量。

- ① 定义 wuteri 为因变量。在源变量表中，选择 wuteri 变量进入【因变量】框。
- ② 定义 mouse 和 etrogen 变量为固定因素变量，选择并送入【固定因子】框。
- (4) 单击【模型】按钮，打开【单变量：模型】对话框，选择【设定】，即自定义模型。
  - ① 在【构建项】栏内的参数框中选择【主效应】项，定义主效应。
  - ② 从【因子与协变量】框中分别选定【mouse】、【etrogen】并移入【模型】框中。
- (5) 在主对话框中单击【确定】按钮，执行多元方差分析过程。输出结果见表 9-18 和表 9-19。

2) 输出结果解释

表 9-18 所示为变量信息，大白鼠子宫重量按大白鼠种系和雌激素剂量分组。因素变量有：种系 mouse，取值 1~4，是种系 A~D 的代码；雌激素剂量 etrogen，取值 1~3，是剂量 0.2、0.4、0.8 的代码。N 是每一单元的样本含量。

表 9-19 所示是方差分析表，在表的左上方标明研究的对象即因变量是子宫重量 wuteri。

表 9-18 因素变量表

		值标签	N
大白鼠种系	1	A	3
	2	B	3
	3	C	3
	4	D	3
雌激素剂量	1	0.2	4
	2	0.4	4
	3	0.8	4

表 9-19 主效应方差分析检验结果

因变量: 子宫重量

源	III 型平方和	df	均方	F	Sig.
校正模型	12531.667 <sup>a</sup>	5	2506.333	27.677	.000
截距	100467.000	1	100467.000	1109.452	.000
mouse	6457.667	3	2152.556	23.771	.001
etrogen	6074.000	2	3037.000	33.537	.001
误差	543.333	6	90.556		
总计	113542.000	12			
校正的总计	13075.000	11			

a. R 方 = .958 (调整 R 方 = .924)

① “源”列。表明偏差来源。这一列表明此列右面将按以下各项列出各统计量：

- “校正模型”。校正模型的第Ⅲ型偏差平方和，即经均值校正后的偏差平方和。在【单变量：模型】对话框中设置的方差分析模型只有两个主效应。该值等于两个主效应 mouse、etrogen 偏差平方和之和。
- “截距”。截距的偏差平方和。
- 主效应 “mouse”。其偏差平方和表明的是由于大白鼠种系不同(对雌激素反应不同)造成的子宫重量之差异，与 etrogen 偏差平方和一样，均属于组间偏差平方和。
- 主效应 “etrogen”。其偏差平方和解释的是不同雌激素剂量造成的子宫重量之差异。
- “误差”。它是除去模型中指定的效应外不可解释的部分。一般情况下，可能包括未考虑到的协变量效应或交互效应、随机因素效应和组内差异。在本例中，其偏差平方和反映组内(即个体之间的)差异，也称为组内偏差平方和。误差项用于检验各效应的假设。其均方值作为 F 检验计算  $F$  值的分母。
- “总计”。是因变量的总偏差平方和在数值上等于截距、两个主效应和误差的偏差平方和之总和。反映因变量原始的总变异。
- “校正的总计”。校正的总偏差平方和。

对方差模型来说，从其值等于校正模型偏差平方和与误差之偏差平方和之总和可以看出，方差模型的总偏差平方和，分解为两个主效应(组间)偏差平方和与误差(组内)偏差平方和。

对于以 wuteri 为因变量，mouse、etrogen 为自变量的线性回归模型来说，校正总计就是线性模型的总偏差平方和，在数值上等于回归平方和与残差平方和之总和。

- ② “Ⅲ型平方和”列。源中所列各项的第Ⅲ类偏差平方和。
- ③ “df”列。源中所列各项的自由度。
- ④ “均方”列。均方在数值上等于偏差平方和除以相应的自由度。
- ⑤ “F”列。即  $F$  值，是各效应项的均方与误差项的均方之比。
- ⑥ “Sig.”列。是进行 F 检验的  $p$  值。

从两个主效应的 F 检验结果的  $p$  值看， $p<0.05$ ，由此得出种系 mouse 和剂量 etrogen 对因变量 wuteri 在 0.05 水平上是有显著性差异的。截距的检验结果  $p<0.05$ ，结论是：对相同剂量的雌激素，不同种系大白鼠子宫重量增加明显不同。因数据较少，可从原始数据观察。

对同种系大白鼠，随雌激素剂量增加，子宫重量增加， $p<0.05$  均值差异显著。

在 wuteri 因变量与 mouse、etrogen 两个自变量之间存在线性回归关系。

3) 应注意的问题

本例中虽然有两个因素变量，但是两个因素变量的各水平构成的每个组合只有一个观测。

故这种试验设计是最简单的双因素方差分析的试验设计方案。因此不能分析因素间的交互作用,无法计算差异的显著性,输出结果也不能给出  $F$  值及其概率。本例按照双因素设计进行方差分析,在不考虑交互作用时会得出较满意的结果。因此,一定要使用模型选项,在构建项中只选择主效应项,而不要选交互项,即不要指定交互项。

### 9.3.4 $2 \times 2$ 析因试验方差分析实例

**【例 5】** 本例使用两种药物 A 和 B 治疗缺铁性贫血病人的数据,是一个  $2 \times 2$  析因试验设计的例题,主要说明均值对比的选项与结果。研究 A、B 两种治疗缺铁性贫血药物的疗效,随机选取 12 个病人分为 4 组,给以不同的治疗:第一组使用一般疗法;第二组使用一般疗法外加药物 A;第三组使用一般疗法外加药物 B;第四组使用一般疗法外加用药物 A 和药物 B。一个月后观察红细胞增加数( $\times 10^6/\text{mm}^3$ ),作析因分析。数据文件为 data09-04。

#### 1) 数据说明与假设

因素变量有两个: drugA 和 drugB, 两个变量均有两个水平,“0”表示不用此药,标签为“no”;“1”表示使用此药,标签为“yes”。因变量为 redcell(红细胞增加数),单位为 $\times 10^6/\text{mm}^3$ 。

该研究的检验假设是  $H_0$ : 药物 A 和药物 B 对患者红细胞增加无显著效果。两种药物无协同作用(即无交互效应)。

#### 2) 操作步骤

(1) 读取数据文件 data09-04。按【分析→一般线性模型→单变量】顺序打开【单变量】主对话框。

(2) 指定分析变量。将变量 redcell 移入【因变量】框。将 drugA 和 drugB 变量移进【固定因子】框,作为因素变量。

(3) 由于本次分析为全模型,因此不用对【单变量:模型】对话框作任何操作。全模型即模型中包括所有主效应和交互效应。对于双因素的全模型应该包括两个主效应 drugA、drugB,一个交互效应 drugA\*drugB。

(4) 在主对话框中,单击【绘制】按钮,打开相应的对话框,要求作 3 个图的操作如下:

① 在【因子】框中选择 drugA,送入【水平轴】栏。单击【添加】按钮,在【图】栏中出现图形表达式 drugA。

② 在【因子】框中选择 drugB,送入【水平轴】栏。单击【添加】按钮,在【图】栏中出现图形表达式 drugB。

③ 在【因子】框中选择 drugA,送入【水平轴】栏,在【因子】框中选择 drugB,作为分线变量送入【单图】栏。单击【添加】按钮,在【图】栏中出现图形表达式 drugA\*drugB。

④ 单击【继续】按钮,返回主对话框。

(5) 主对话框中,单击【选项】按钮,打开相应的对话框。

① 在【因子与因子交互】框中分别选择因素变量 drugA、drugB、drugA\*drugB 和 (Overall),单击向右箭头按钮,将它们送入【显示均值】框中。单击【继续】按钮回到主对话框。

② 在【输出】栏内选择【描述统计】,见图 9-15。

(6) 单击【确定】按钮,提交系统执行。

#### 3) 输出结果

输出结果见表 9-20~表 9-26、图 9-17~图 9-19。

4) 结果说明与分析

表 9-20 所示是“两种药物对红细胞增加数作用的研究”课题中的变量信息。表中列出了 drugA 和 drugB 两个因素变量和分类水平，以及每个水平的样本含量。

表 9-21 所示为描述统计量。

表 9-22 所示为方差分析结果。可以看出：

- ① 总校正偏差平方和分解为校正模型的偏差平方和与随机误差的偏差平方和。校正模型的偏差平方和=drugA 偏差平方和+drugB 偏差平方和+交互效应 drugA\*drugB 的偏差平方和。
- ② 随机误差偏差平方和为 0.08。
- ③ 各项偏差平方和除以各自的自由度是相应的均方。各项  $F$  值为各项均方除以误差均方。
- ④  $F$  检验的结果，显著性概率  $p$  值均小于 0.01。

表 9-20 研究中的变量信息

主体间因子			
		值标签	N
A药	0	no	6
	1	yes	6
B药	0	no	6
	1	yes	6

表 9-21 描述统计量

因变量: 红细胞增加量				
A药	B药	均值	标准 偏差	N
0 no	0 no	.800	.1000	3
	1 yes	1.000	.1000	3
	总计	.900	.1414	6
1 yes	0 no	1.200	.1000	3
	1 yes	2.100	.1000	3
	总计	1.650	.5010	6
总计	0 no	1.000	.2366	6
	1 yes	1.550	.6091	6
	总计	1.275	.5259	12

表 9-22 方差分析表

因变量: 红细胞增加量					
源	III 型平方和	df	均方	F	Sig.
校正模型	2.963 <sup>a</sup>	3	.988	98.750	.000
截距	19.508	1	19.508	1950.750	.000
drugA	1.687	1	1.687	168.750	.000
drugB	.908	1	.908	90.750	.000
drugA * drugB	.368	1	.368	36.750	.000
误差	.080	8	.010		
总计	22.550	12			
校正的总计	3.043	11			

a. R 方 = .974 (调整 R 方 = .964)

结论: drugA、drugB 均对红细胞的增加有显著疗效。并且交互效应也很显著。检验结果拒绝无效假设,使用药物 A 与不使用药物 A 的红细胞增加数的均值有显著性差异。使用药物 B 与不使用药物 B 的红细胞增加数的均值有显著性差异。同时使用药物 A 和药物 B,两药物协同作用也很显著。

表 9-23~表 9-26 为红细胞增加数的估计的边际值表。总结这 4 个表成为表 9-27。可以看出,交互项产生单元格中均数,主效应项生成边际均数,总效应项 (OVERALL) 生成总均数 1.275。

表 9-23 drugA 边际均值估计值表

2. A药					
因变量: 红细胞增加量					
A药	均值	标准 误差	95% 置信区间		
			下限	上限	
0 no	.900	.041	.806	.994	
1 yes	1.650	.041	1.556	1.744	

表 9-24 drugB 边际均值估计值表

3. B药					
因变量: 红细胞增加量					
B药	均值	标准 误差	95% 置信区间		
			下限	上限	
0 no	1.000	.041	.906	1.094	
1 yes	1.550	.041	1.456	1.644	

表 9-25 交互项边际均值估计值表

4. A药 * B药					
因变量: 红细胞增加量					
A药	B药	均值	标准 误差	95% 置信区间	
				下限	上限
0 no	0 no	.800	.058	.667	.933
	1 yes	1.000	.058	.867	1.133
1 yes	0 no	1.200	.058	1.067	1.333
	1 yes	2.100	.058	1.967	2.233

表 9-26 综合边际均值估计值表

1. 总均值			
因变量: 红细胞增加量			
均值	标准 误差	95% 置信区间	
		下限	上限
1.275	.029	1.208	1.342

表 9-27 边际值估计值示意

实验分组		A 药		B 边际均值
		不用	使用	
B 药	不用	0.80	1.20	1.000
	使用	1.00	2.10	1.550
A 边际均值		0.900	1.650	1.275

根据 drugA、drugB 使用【转换】菜单中的第一项【计算变量】功能生成有 4 个水平的新变量，分别代表一般治疗、一般治疗加 A 药、一般治疗加 B 药和一般治疗加 A 药和 B 药，利用多重比较功能比较 4 种用药方法的疗效。

图 9-17、图 9-18 和图 9-19 是一系列边际均值图。读者可以对照表 9-27 查看数据与图的关系。从图 9-17 和图 9-18 可以看出，每种药物单独效应：用药与不用药对红细胞增加数的效用。在图 9-19 中可以看出两直线明显不平行，因此很明显，这两种药之间存在交互效应。

如果想更直观地比较 4 种疗法的疗效，可以

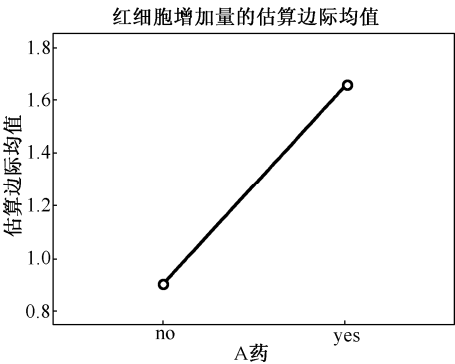


图 9-17 A 药效应红细胞增加数均值图

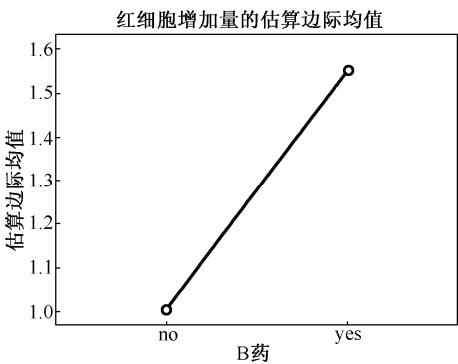


图 9-18 B 药效应红细胞增加数均值图

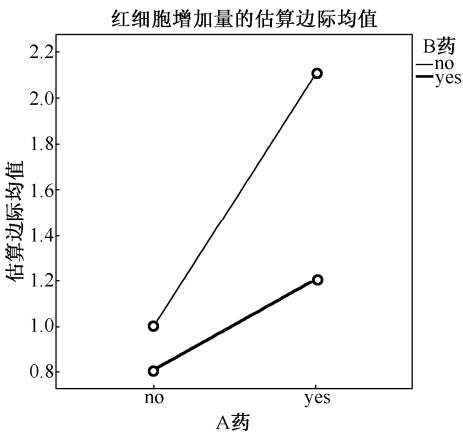


图 9-19 A、B 药对红细胞增加数交互效应边际图

9.3.5 拉丁方区组设计的方差分析实例

**【例 6】** 拉丁方试验设计的特点是有两个以上因素变量，每个因素变量的水平数相等。

变量：variety(甜菜种系)、rep(地块行)、col(地块列)harvest(收获次数)、yield(产量)。要求分析 6 种甜菜品种在相同土壤条件下的产量是否有显著性差异。为了得出这一结论，同时检验地块是否对平均产量有影响，即地块的行与行之间、列与列之间的平均产量是否有显著性差异，将 6 种甜菜种子播在 6 行 6 列的地块上，记录两次收获的产量。数据文件为 data09-05。试验的假设是：不同地块(行、列)对产量均值无影响，不同种子产量均值间也无显著差异。

1) 操作步骤

分两步完成，先作方差分析，再作边际值估计值表。

(1) 读取数据文件 data09-05。按【分析→一般线性模型→单变量】顺序单击菜单项，最后打开【单变量】主对话框。

(2) 在主对话框中定义分析变量。

① 将 yield 变量移入【因变量】框。

② 将 rep、col、variety 变量进入【固定因子】框，这些变量作为因素变量。

(3) 在主对话框中，单击【模型】按钮，打开相应的对话框。在该对话框中选择【设定】，自定义模型：指定要求分析 3 个主效应 rep、col、variety。单击【继续】按钮，返回主对话框。

(4) 在主对话框中，单击【选项】按钮，打开相应的对话框。选择 3 个因素变量 rep、col、variety 和 (Overall) 送入【显示均值】栏内，选择【比较主效应】，其他使用默认值。

(5) 在主对话框中单击【确定】按钮，提交系统执行，完成方差分析。

2) 输出结果(见表 9-28~表 9-32)

表 9-28 所示为方差分析表，只对 rep、col、variety 变量的主效应作方差分析。方差分析解决 3 个因素变量的各水平，产量平均值之间差异是否有统计意义。

表 9-28 方差分析表

因变量: 产量

源	III 型平方和	df	均方	F	Sig.
校正模型	27.717 <sup>a</sup>	15	1.848	1.339	.211
截距	22588.751	1	22588.751	16364.072	.000
rep	4.460	5	.892	.646	.666
col	1.695	5	.339	.246	.940
variety	21.563	5	4.313	3.124	.015
误差	77.302	56	1.380		
总计	22693.770	72			
校正的总计	105.019	71			

a. R 方 = .264 (调整 R 方 = .067)

表 9-29 各列、各行、各种甜菜产量的分类均值表(边际均值)

因变量: 产量

行号	均值	标准 误差	95% 置信区间	
			下限	上限
1	17.850	.339	17.171	18.529
2	17.658	.339	16.979	18.338
3	18.017	.339	17.337	18.696
4	17.933	.339	17.254	18.613
5	17.517	.339	16.837	18.196
6	17.300	.339	16.621	17.979

因变量: 产量

列号	均值	标准 误差	95% 置信区间	
			下限	上限
1	17.483	.339	16.804	18.163
2	17.650	.339	16.971	18.329
3	17.642	.339	16.962	18.321
4	17.942	.339	17.262	18.621
5	17.875	.339	17.196	18.554
6	17.683	.339	17.004	18.363

因变量: 产量

甜菜种系编号	均值	标准 误差	95% 置信区间	
			下限	上限
1	17.367	.339	16.687	18.046
2	17.817	.339	17.137	18.496
3	17.475	.339	16.796	18.154
4	17.367	.339	16.687	18.046
5	18.883	.339	18.204	19.563
6	17.367	.339	16.687	18.046

查看各主效应的 Sig. 值，只有因素变量 variety 的值为 0.015，小于 0.05。可得出结论：6 种甜菜的平均产量具有显著性差异。平均产量的差异主要是品种不同造成的。

表 9-30 主效应因素均值表(rep、col、variety)

因变量: 产量					因变量: 产量											
(I) 行号	(J) 行号	均值差值 (I-J)	标准 误差	Sig. <sup>a</sup>	差分的 95% 置信区间 <sup>a</sup>		(I) 列号	(J) 列号	均值差值 (I-J)	标准 误差	Sig. <sup>a</sup>	差分的 95% 置信区间 <sup>a</sup>				
					下限	上限						下限	上限			
1	2	.192	.480	.691	-.769	1.153	1	2	-.167	.480	.730	-1.128	.794			
	3	-.167	.480	.730	-1.128	.794		1	3	-.158	.480	.743	-1.119	.803		
	4	-.083	.480	.863	-1.044	.878			1	4	-.458	.480	.343	-1.419	.503	
	5	.333	.480	.490	-.628	1.294				1	5	-.392	.480	.418	-1.353	.569
	6	.550	.480	.256	-.411	1.511					1	6	-.200	.480	.678	-1.161
2	1	-.192	.480	.691	-1.153	.769	2					1	.167	.480	.730	-.794
	3	-.358	.480	.458	-1.319	.603		2				3	.008	.480	.986	-.953
	4	-.275	.480	.569	-1.236	.686			2			4	-.292	.480	.546	-1.253
	5	.142	.480	.769	-.819	1.103				2		5	-.225	.480	.641	-1.186
	6	.358	.480	.458	-.603	1.319					2	6	-.033	.480	.945	-.994
3	1	.167	.480	.730	-.794	1.128	3					1	.158	.480	.743	-.803
	2	.358	.480	.458	-.603	1.319		3				2	-.008	.480	.986	-.969
	4	.083	.480	.863	-.878	1.044			3			4	-.300	.480	.534	-1.261
	5	.500	.480	.302	-.461	1.461				3		5	-.233	.480	.629	-1.194
	6	.717	.480	.141	-.244	1.678					3	6	-.042	.480	.931	-1.003
4	1	.083	.480	.863	-.878	1.044	4					1	.458	.480	.343	-.503
	2	.275	.480	.569	-.686	1.236		4				2	.292	.480	.546	-.669
	3	-.083	.480	.863	-1.044	.878			4			3	.300	.480	.534	-.661
	5	.417	.480	.389	-.544	1.378				4		5	.067	.480	.890	-.894
	6	.633	.480	.192	-.328	1.594					4	6	.258	.480	.592	-.703
5	1	-.333	.480	.490	-1.294	.628	5					1	.392	.480	.418	-.569
	2	-.142	.480	.769	-1.103	.819		5				2	.225	.480	.641	-.736
	3	-.500	.480	.302	-1.461	.461			5			3	.233	.480	.629	-.728
	4	-.417	.480	.389	-1.378	.544				5		4	-.067	.480	.890	-1.028
	6	.217	.480	.653	-.744	1.178					5	6	.192	.480	.691	-.769
6	1	-.550	.480	.256	-1.511	.411	6					1	.200	.480	.678	-.761
	2	-.358	.480	.458	-1.319	.603		6				2	.033	.480	.945	-.928
	3	-.717	.480	.141	-1.678	.244			6			3	.042	.480	.931	-.919
	4	-.633	.480	.192	-1.594	.328				6		4	-.258	.480	.592	-1.219
	5	-.217	.480	.653	-1.178	.744					6	5	-.192	.480	.691	-1.153

基于估算边际均值

a. 对多个比较的调整：最不显著差别（相当于未作调整）。

基于估算边际均值

a. 对多个比较的调整：最不显著差别（相当于未作调整）。

因变量: 产量		均值差值 (I-J)	标准 误差	Sig. <sup>b</sup>	差分的 95% 置信区间 <sup>b</sup>	
(I) 甜菜种系编号	(J) 甜菜种系编号				下限	上限
1	2	-.450	.480	.352	-1.411	.511
	3	-.108	.480	.822	-1.069	.853
	4	2.331E-015	.480	1.000	-.961	.961
	5	-1.517 <sup>*</sup>	.480	.003	-2.478	-.556
	6	2.121E-014	.480	1.000	-.961	.961
2	1	.450	.480	.352	-.511	1.411
	3	.342	.480	.479	-.619	1.303
	4	.450	.480	.352	-.511	1.411
	5	-1.067 <sup>*</sup>	.480	.030	-2.028	-.106
	6	.450	.480	.352	-.511	1.411
3	1	.108	.480	.822	-.853	1.069
	2	-.342	.480	.479	-1.303	.619
	4	.108	.480	.822	-.853	1.069
	5	-1.408 <sup>*</sup>	.480	.005	-2.369	-.447
	6	.108	.480	.822	-.853	1.069
4	1	-2.331E-015	.480	1.000	-.961	.961
	2	-.450	.480	.352	-1.411	.511
	3	-.108	.480	.822	-1.069	.853
	5	-1.517 <sup>*</sup>	.480	.003	-2.478	-.556
	6	1.887E-014	.480	1.000	-.961	.961
5	1	1.517 <sup>*</sup>	.480	.003	.556	2.478
	2	1.067 <sup>*</sup>	.480	.030	.106	2.028
	3	1.408 <sup>*</sup>	.480	.005	.447	2.369
	4	1.517 <sup>*</sup>	.480	.003	.556	2.478
	6	1.517 <sup>*</sup>	.480	.003	.556	2.478
6	1	-2.121E-014	.480	1.000	-.961	.961
	2	-.450	.480	.352	-1.411	.511
	3	-.108	.480	.822	-1.069	.853
	4	-1.887E-014	.480	1.000	-.961	.961
	5	-1.517 <sup>*</sup>	.480	.003	-2.478	-.556

基于估算边际均值

\*. 均值差值在 .05 级别上较显著。

b. 对多个比较的调整：最不显著差别（相当于未作调整）。



表 9-31 单变量方差分析的 3 个表(rep、col、variety)

因变量: 产量						因变量: 产量					
	平方和	df	均方	F	Sig.		平方和	df	均方	F	Sig.
对比	4.460	5	.892	.646	.666	对比	1.695	5	.339	.246	.940
误差	77.302	56	1.380			误差	77.302	56	1.380		

F 检验 行号 的效应。该检验基于估算边际均值间的线性独立成对比较。

因变量: 产量					
	平方和	df	均方	F	Sig.
对比	21.563	5	4.313	3.124	.015
误差	77.302	56	1.380		

F 检验 甜菜种系编号 的效应。该检验基于估算边际均值间的线性独立成对比较。

表 9-29 所示为各列、行和各个品种的边际均值估计值表，此外还有标准误和区间估计。

表 9-30 包括 3 个表，为每个因素的各水平均值的成对比较表。每个表中给出各变量两两水平之间的均值之差、均值差的标准误、针对两均值相等的假设检验的显著性概率 Sig. 值、差值的 95% 置信区间。从 3 个表中可以看到，只有第 5 种种子比其他 5 种种子产量都高，且差值具有明显的统计意义。

表 9-31 包括 3 个表，为各因素单变量方差分析表。表中给出  $F$  值及大于等于该值的概率。可以看出，只有种类的方差分析的 Sig. 值为 0.015，小于 0.05。

综上所述，产量主要受种子的影响，而第 5 种种子的产量明显高于其他种子；产量与地块所处位置行、列无关。

表 9-32 是最后给出的综合统计表，给出产量的总均值、均值标准误和 95% 置信区间。

本例中虽然有 3 个因素变量，但 3 个因素变量的各水平组合构成的每个单元只有 1 个观测。实际上，这种试验设计如果分析因素间的交互作用，无法计算差异的显著性，因此输出结果不能给出大于等于  $F$  值的概率。如果本例按照三因素设计进行方差分析，不考虑交互作用会得出较满意的结果。这就需要注意，一定要使用模型选项，在构建项里只选择主效应项，而不要选择交互项。所以在进行方差分析时不考虑交互作用，而只考虑主效应，要求边际均值时才用全模型。如果考虑两次收获(变量 harvest)，则行、列、地块的每种组合中有两次收获的数据，那么就可以考虑交互作用了。另外，该试验中，行、列两个因素变量不是相互独立的，因此不是一个很严格的设计，在这里仅为说明拉丁方设计的概念及其解决该问题而使用 SPSS 软件的方法。

表 9-32 总均值表

因变量: 产量			
均值	标准 误差	95% 置信区间	
		下限	上限
17.713	.138	17.435	17.990

9.3.6 协方差分析实例

协方差分析是利用线性回归方法消除混杂因素的影响后进行的方差分析，也就是先从因变量的总偏差平方和中去掉协变量对因变量的回归平方和，再对残差平方和进行分解，进行方差分析。例如，考虑药物对患者某个生化指标变化的影响，要比较试验组与对照组该指标的变化均值是否有显著性差异，以确定药物的有效性。可能要考虑患者病程的长短、年龄以及原指标水平对疗效的影响。消除这些因素的影响，考虑药物疗效，才是科学的分析方法。有些试验可

以考虑观测对象的选择,使这些条件都一致。例如,选择同品种、同一胎的大白鼠分组,在相同的饲养条件下进行试验,可以相应地避免许多混杂因素的影响。其他试验很难避免,因此要考虑使用协方差分析方法。这些混杂因素变量称作协变量。

协方差分析中要求因变量应该是等间隔测量的变量,理论上要求其服从正态分布。因素变量是分类变量,并且相互独立。协变量是与因变量存在一定相关关系(相互不独立)的等间隔测量的变量。因变量与协变量之间是否线性相关,可以通过经验得知或使用【图形】菜单中作散点图的功能进行初步的直观判断。

**【例 7】** 数据文件 data09-06 是镉作业工人年龄与肺活量的数据,数据来源于《医用统计方法》(金丕焕,人民卫生出版社)。镉作业工人按暴露于镉粉尘的年数分为大于等于 10 年和不足 10 年两组。两组工人的年龄未经控制(人随着年龄的增长,肺活量也会有所下降),测量了每个工人的肺活量。课题研究暴露于镉粉尘中的年数与肺活量的关系。数据变量如下:time(接触镉粉尘时间分组),取值 1 代表大于等于 10 年,2 代表不足 10 年;age(年龄);vitalcp(肺活量,单位为 L)。

1) 操作步骤

(1) 读取数据文件 data09-06。按【分析→一般线性模型→单变量】顺序单击菜单项,打开【单变量】主对话框。

(2) 指定分析变量。将肺活量 vitalcp 变量移入【因变量】框;将暴露时间分组 time 变量送入【固定因子】框,time 变量作为因素变量;将年龄 age 变量送入【协变量】框,即 age 变量作为协变量。

(3) 在主对话框中,单击【选项】按钮,打开相应的对话框。

① 在【因子与因子交互】框中选择因素变量 time,将其送入【显示均值】框。要求输出暴露于镉粉尘年数大于等于 10 年、不足 10 年两组工人的肺活量平均值。

② 在【输出】栏内选中【参数估计】,要求输出年龄作自变量,肺活量作因变量的线性回归方程的参数。

(4) 单击【继续】按钮,返回主对话框。单击【确定】按钮,提交系统执行。

2) 输出结果(见表 9-33~表 9-36)

表 9-33 所示为因素变量表,列出了按时间分组的变量标签、样本量。

表 9-34 所示为方差分析结果。

表 9-33 因素变量表

	值标签	N
暴露时间 1	>=10年	12
2	<10年	16

表 9-34 方差分析表

因变量: 肺活量					
源	III 型平方和	df	均方	F	Sig.
校正模型	11.085 <sup>a</sup>	2	5.543	10.073	.001
截距	41.936	1	41.936	76.216	.000
age	10.881	1	10.881	19.775	.000
time	.542	1	.542	.985	.330
误差	13.755	25	.550		
总计	483.625	28			
校正的总计	24.841	27			

a. R 方 = .446 (调整 R 方 = .402)

- ① 表的左上方列出因变量为“肺活量”,即研究对象。
- ② 分别列出方差来源,系统默认的第III类型的偏差平方和、自由度、均方差、F 值和 Sig.。

③ 从方差分析表中看到，总的偏差平方和 24.841 被分解为条件引起的平方和(校正模型)11.085 和试验误差引起的平方和 13.755。从显著性概率(Sig.)看,time 的概率为 0.330,大于 0.05。协变量效应由协变量 age 决定，其偏差平方和为 10.881,  $p = 0.000$ ，小于 0.001。因此可以得出结论：肺活量的差异是由于被试者的年龄差异所致，与被试者接触镉粉尘的时间是否大于 10 年无关。

表 9-35 所示是在【单变量：选项】对话框中的【输出】栏内选中了【参数估计】的输出结果。这里主要给出了 age 作为自变量、vitalcp 作为因变量的线性回归方程的斜率，即变量 age 的回归系数值为-0.087。这一回归系数也是符合生理常识的。因为成年人随着年龄的增长，肺活量会有所下降。

表 9-36 所示是在【单变量：选项】对话框中，将 time 移入【显示均值】框的结果。表中按 time 分组分别列出了平均值、标准误和 95%的置信区间。因素变量各单元均值是 10 年以下组的肺活量均值为 4.219, 10 年以上组的肺活量均值为 3.919。协方差分析结果表明，这两组肺活量均值差异无统计意义上的显著性。

表 9-35 参数估测值的输出结果表

因变量: 肺活量						
参数	B	标准 误差	t	Sig.	95% 置信区间	
					下限	上限
截距	7.977	.886	8.998	.000	6.151	9.803
age	-.087	.020	-4.447	.000	-.127	-.047
[time=1]	.300	.303	.993	.330	-.323	.924
[time=2]	0 <sup>a</sup>	.	.	.	.	.

a. 此参数为冗余参数，将被设为零。

表 9-36 按时间分组的肺活量均值表

因变量: 肺活量				
暴露时间	均值	标准 误差	95% 置信区间	
			下限	上限
1 >=10年	4.219 <sup>a</sup>	.223	3.761	4.678
2 <10年	3.919 <sup>a</sup>	.191	3.526	4.312

a. 模型中出现的协变量在下列值处进行评估: 年龄 = 46.64.

9.3.7 多维交互效应方差分析实例

【例 8】 本例主要表明使用单变量过程进行多因素方差分析构成模型的灵活性。

1) 试验数据

本例为教育心理学试验，数据是心理运动测验分数与被试者必须瞄准的目标大小关系的资料。

- (1) 选择 4 个大小不同的目标(target)：1(T1)、2(T2)、3(T3)、4(T4)。
- (2) 从若干使用过的设备中选择 3 部测验设备(device)：1(D1)、2(D2)、3(D3)。
- (3) 选择 2 种不同明暗程度的照明环境(light)：1(L1)、2(L2)。

4 个大小不同的目标、3 部设备、2 种不同的照明环境构成 4×3×2 的析因试验设计。不同目标、设备与照明水平构成了 24 个组合的单元。每一个组合中随机部署 5 名被试者进行测试心理运动得分，得到 120 个得分数据。

每个观测为被试者在同一条件组合下的 5 个得分。数据文件为 data09-07。

2) 操作步骤

(1) 读取数据文件 data09-07。按【分析→一般线性模型→单变量】顺序单击菜单项，打开【单变量】主对话框。

(2) 将 score 变量移入【因变量】框，即测试得分为因变量；将 target、device、light 变量移入【固定因子】框，作为因素变量。

(3) 在主对话框中，单击【模型】按钮，打开相应的对话框。首先选择【设定】项，自定义模型，激活对话框中的各控制功能。

① 在【构建项】栏中的参数框内选择【主效应】项。在【因子与协变量】框中选择 target、device、light 变量进入【模型】框中，即将这 3 个作为主效应定义到模型中。

② 选择交互项。在【构建项】栏中的参数框内选择交互项。在【因子与协变量】框中，选择一个变量 target，按住 Ctrl 键，选择第二个变量 device，按右移箭头按钮，将 target\*device 交互项移入到【模型】框中，即该交互项进入模型。再用同样方法在模型中建立另一个二次交互项 target\*light。

与建立二次交互项同样的方法在模型中建立三次交互项 target\*device\*light。

(4) 在主对话框中，单击【选项】按钮，打开【选项】对话框。在【因子与因子交互】框中指定 target\*device\*light，将其送入【显示均值】框，目的是要输出各单元格的均值。

(5) 在主对话框中，单击【绘制】按钮，打开相应对话框，选择 target 变量作水平轴变量，选择 device 变量作为分线变量，选择 light 变量作为分图变量，分别送入右边的 3 个栏中，然后单击【添加】按钮，将作图表达式 target\*device\*light 送入【图】框中。

(6) 单击【确定】按钮，执行运算。

3) 运行结果(见表 9-37～表 9-39 和图 9-21、图 9-22)

4) 输出结果解释与分析

表 9-37 所示为原始数据综合信息：系统接受了 120 个观测；列出各个因素变量，变量值标签和样本含量。

表 9-38 为方差分析表，表的左上方标有因变量“得分”(score)。可以看出，对于模型来说，校正模型与截距的平方和加上误差的平方和等于总平方和。从方差分析的角度来看，各主效应的平方和与各交互效应的平方和以及误差的平方和，其总和就是校正的总计。而误差是校正的总计平方和与校正模型的平方和之差。也就是说，如果模型包括表中这些主效应和交互效应，那么它与校正模型之间平方和之差是 70.4。

从表 9-38 中的 Sig. 列数值还可以看出，该试验中各主效应和交互效应都是有统计意义的。

表 9-37 因素变量表

		值标签	N
目标	1	t1	30
	2	t2	30
	3	t3	30
	4	t4	30
设备	1	d1	40
	2	d2	40
	3	d3	40
亮度	1	l1	60
	2	l2	60

表 9-38 方差分析结果

因变量: 得分					
源	III 型平方和	df	均方	F	Sig.
校正模型	783.467 <sup>a</sup>	23	34.064	46.451	.000
截距	3162.133	1	3162.133	4312.000	.000
target	235.200	3	78.400	106.909	.000
device	86.467	2	43.233	58.955	.000
light	76.800	1	76.800	104.727	.000
target * light	93.867	3	31.289	42.667	.000
target * device	104.200	6	17.367	23.682	.000
target * device * light	186.933	8	23.367	31.864	.000
误差	70.400	96	.733		
总计	4016.000	120			
校正的总计	853.867	119			

a. R 方 = .918 (调整 R 方 = .898)

表 9-39 列出了 3 个因素变量构成的单元格表，给出了各单元格的均值、标准误和 95% 的置信区间。从均值列数据可以看出，light=1、devise=2、target=3 这个条件组合的测试平均分最高，light=1、devise=2、target=1 的条件组合的测试平均分最低。心理学专业人士可以根据均值表和方差分析表得出专业性的结论。

图 9-20 和图 9-21 更直观地表现了表 9-39 中“均值”栏中的数据，而且可以很清楚地看出在不同的光照条件下目标变量与设备之间均存在交互效应。读者可以自己作出使用不同设备时，光照与目标之间的边际均值图。

表 9-39 各单元格观测均值

因变量: 得分			均值	标准 误差	95% 置信区间	
设备	亮度	目标			下限	上限
1 d1	1 l1	1 t1	2.000	.383	1.240	2.760
		2 t2	8.000	.383	7.240	8.760
		3 t3	8.800	.383	8.040	9.560
		4 t4	7.200	.383	6.440	7.960
	2 l2	1 t1	1.600	.383	.840	2.360
		2 t2	4.400	.383	3.640	5.160
		3 t3	5.600	.383	4.840	6.360
		4 t4	6.800	.383	6.040	7.560
2 d2	1 l1	1 t1	.800	.383	.040	1.560
		2 t2	8.400	.383	7.640	9.160
		3 t3	9.600	.383	8.840	10.360
		4 t4	9.200	.383	8.440	9.960
	2 l2	1 t1	5.200	.383	4.440	5.960
		2 t2	6.400	.383	5.640	7.160
		3 t3	4.800	.383	4.040	5.560
		4 t4	2.800	.383	2.040	3.560
3 d3	1 l1	1 t1	4.000	.383	3.240	4.760
		2 t2	5.200	.383	4.440	5.960
		3 t3	6.000	.383	5.240	6.760
		4 t4	2.000	.383	1.240	2.760
	2 l2	1 t1	3.200	.383	2.440	3.960
		2 t2	1.200	.383	.440	1.960
		3 t3	4.400	.383	3.640	5.160
		4 t4	5.600	.383	4.840	6.360

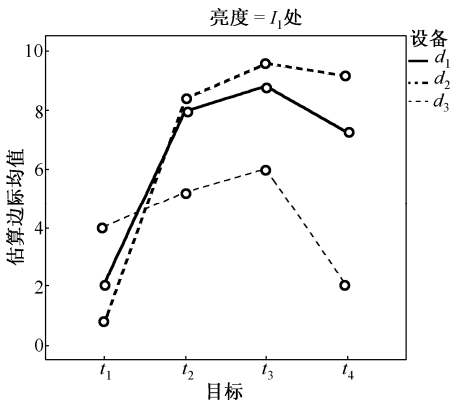


图 9-20 第一种照度下心理得分边际均值图

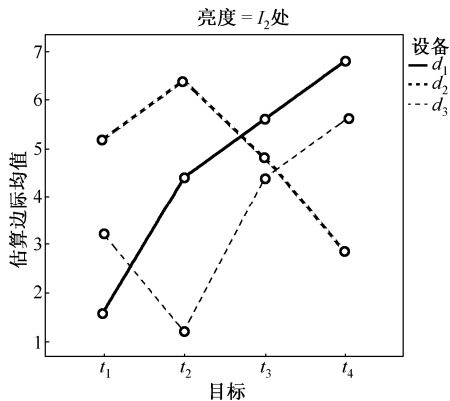


图 9-21 第二种照度下心理得分边际均值图

9.4 多因变量线性模型的方差分析

9.4.1 多因变量方差分析概述

SPSS 的一般线性模型中的多变量过程提供多因变量的方差分析。多因变量方差分析模型的因变量是尺度变量(连续变量)。分类变量作为固定因素变量,协变量必须是尺度变量。该模型是基于尺度因变量与作为预测因子的因素变量和协变量之间的相关关系。一般线性模型中的多变量过程构造的模型是一般线性模型。可以检验因变量在因素变量各水平组合中的组均值的

效应,可以研究因素间的交互效应和单一因素的效应,另外还包括协变量效应和协变量与因素间的交互效应。对于回归分析,协变量作为自变量即预测变量。

一般线性模型中的多变量过程可以检验平衡和不平衡模型。模型中每个单元包括相同数量的观测为平衡设计。在多因变量模型中,模型中的效应平方和和误差平方和是矩阵形式的,而不是像在单因变量模型中的格式,这些矩阵称作 **SSCP** 矩阵(平方和与叉积矩阵)如果指定了不止一个因变量,多因素方差分析使用 **Pillai** 迹、**Wilks**、**Hotelling** 迹和 **Roy** 最大根判据和近似 **F** 统计量以及对每个因变量的单变量方差分析。除检验假设外,一般线性模型多因变量分析还产生参数估计。

通常使用 **Priori** 对比执行假设检验。当 **F** 检验已经表明显著性后,还可以使用多重比较检验评价指定均值间的差异。对边际均值的估计给出单元格预测均值的估计,这些边际均值图很容易将某些关系可视化。对每个因变量可分别进行多重比较检验。

残差、预测值、**Cook** 距离、杠杆值可以作为新变量保存在数据文件中,以便验证假设。可以求残差的 **SSCP** 矩阵,它是残差平方和与叉积的矩阵,残差协方差矩阵是残差的 **SSCP** 矩阵除以残差自由度。还有残差相关矩阵,这是标准化的残差协方差矩阵。

**WLS** 权重选项允许指定一个变量,给观测不同的权重用于加权最小平方分析,或用作对不同测量精度的补偿。

为了检验有关参数估计的假设,一般线性模型多变量过程假设:模型中的观测和因变量之间的误差值是彼此独立的。一个好的研究设计一般要避免违反这个假设。

因变量的协方差在各单元中是常数。当单元(因素变量水平组合)中数量(所包括的观测数)不同时,这一点尤其重要。

因变量的误差方差在因素变量水平的各组合中是相等的,即误差方差具有齐性。

**SPSS** 一般线性模型多变量过程用分析菜单中的一般线性模型所属的多变量菜单调用。

## 9.4.2 多因变量方差分析过程和数据要求

### 1. 多因变量方差分析的数据

多因变量方差分析的因变量应该是数值型尺度变量,即连续变量。

因素变量是分类变量,可以是数值型的,也可以是变量值小于或等于 8 个字符的字符型的。分类预测因素可以选择作为模型中的自变量。因素的每个水平可以与因变量的值有不同的线性效应。一般线性模型多变量过程假设所有的模型因素都是固定的。也就是说,通常按照设计,它们被认为是所有感兴趣的变量值,都出现在数据文件中。

协变量是与因变量相关的数值型变量。

因变量数据是多元正态分布的随机样本。在总体中,方差-协方差矩阵对所有单元都是相等的。要检验假设,可以用方差齐性检验(包括 **Box's M** 检验)和用分布-水平图,还可以检验残差和作残差图。

在进行方差分析之前,有必要使用探索分析过程探索数据。对单个因变量,使用一般线性模型多变量进行方差分析。如果在不同情况下对每个被试对象测试同一个因变量,可以使用一般线性模型重复度量过程进行重复测量的方差分析。

### 2. 操作方法

按【分析→一般线性模型→多变量】顺序单击菜单项,打开【多变量】主对话框,如图 9-22 所示。

多因变量方差分析过程与单因变量方差分析过程操作相同的有 9.3 节中的模型功能(设计分析模型)、对比功能(选择对照方法)、绘制图形功能(设定边际均值图参数)、两两比较功能(选择多重比较方法)、保存功能(选择要保存的输出变量)。虽然操作相同,输出结果却不同。因为有多多个因变量,对每个因变量有一组输出。例如,在【单变量:轮廓图】对话框中指定了一个三维边际均值图形表达式,则对每个因变量,均按该表达式生成一组边际均值图。

多因变量方差分析过程与单因变量方差分析过程操作上的不同之处在于,前者在【因变量】框中可以选择多个因变量,而后者只能选择一个因变量,【选项】功能也与单因变量方差分析过程略有不同。

在主对话框中,单击【选项】按钮,打开如图 9-23 所示的【多变量:选项】对话框。



图 9-22 【多变量】主对话框

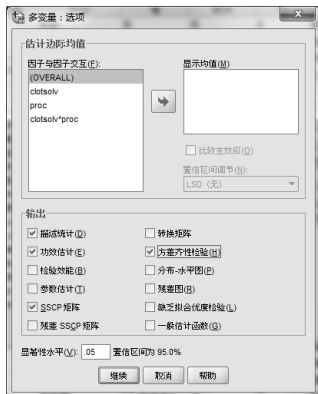


图 9-23 【多变量:选项】对话框

(1) 【估计边际均值】栏中选择要估计的边界均值。操作方法见 9.3.2 节。

(2) 【输出】框。选择输出项。

- 【描述统计】。输出描述统计量,有观测的均值、标准差和每个单元格中的观测数。
- 【功效估计】。输出效应量估计。选择此项,给出  $\eta^2$  (eta square) 值。它是由一个自变量所解释的变异 (SSH) 对自变量解释的变异和未计入模型解释的变异 (SSE) 总和 (SSH+SSE) 之比 (SSH/(SSH+SSE))。
- 【检验效能】。给出 F 检验的概率。它检验的是组间差异,在假设是基于观测值时,检验各种假设的功效。计算功效的显著性水平,系统默认临界值为 0.05。
- 【参数估计】。给出了各因素变量的模型参数估计、标准误、T 检验的  $t$  值、显著性概率和 95% 的置信区间。
- 【SSCP 矩阵】。对每个效应给出平方和与叉积矩阵。对设计中的每个效应给出假设的和误差的 SSCP 矩阵。每个组间效应有不同的 SSCP 矩阵,对所有组间效应只有一个误差矩阵。
- 【残差 SSCP 矩阵】。给出 RSSCP 残差的平方和与叉积矩阵。RSSCP 的维度与模型中因变量数相同。残差的协方差矩阵为 RSSCP 除以残差自由度。残差相关矩阵是由残差协方差矩阵标准化得来的。
- 【转换矩阵】。显示对因变量的转换系数矩阵或 M 矩阵。
- 【方差齐性检验】。给出方差齐性检验结果。Levene 检验每个因变量在所有因素的水平组合间方差是否相等。
- 【分布-水平图】。绘制观测单元均值-标准差图和观测单元均值-方差图。

- **【残差图】**。绘制残差图。给出“观测值\*预测值\*标准化”残差图。
  - **【缺乏拟合优度检验】**(应为**【失拟检验】**)。检查独立变量和非独立变量间的关系是否被充分描述。执行一种拟合不足检验(它要求对一个或几个自变量重复观测)。如果检验被拒绝就意味着当前的模型不能充分说明响应变量与预测因素之间的关系,可能有变量被忽略或模型中需要其他项。
  - **【一般估计函数】**。产生表明估计函数一般形式的表格。可以根据一般估计函数通过 LMATRIX 子命令自定义假设检验。
- (3) 在**【显著性水平】**框中,改变**【置信区间】**框内多重比较的显著性水平。

9.4.3 多因变量线性模型方差分析实例

**【例 9】** 本例数据是 1481 个心梗患者的数据,数据文件为 data09-08。

1) 变量说明(见表 9-40)

作为对心梗(MI 或心脏病发作)的初步治疗,有时在手术之前给溶解血栓(凝块消融药)的药物,帮助清理患者的动脉。3 种可用的药物是 alteplase 阿替普酶、reteplase 瑞替普酶和 streptokinase 链激酶。阿替普酶和瑞替普酶是新药,较昂贵。一个地区的卫生保健系统想确定,是否他们的价格-效应足够代替链激酶。溶解血栓的药物有一个好处就是外科手术比较平稳,因而痊愈周期比较短。如果新药是有效的,患者住院的时间就会较短。地区的卫生保健系统希望,较短的住院时间将有助于补偿术前新药的较大的花费。

表 9-40 变量说明

变量名	标 签	中文标签	值	值标签	值标签(中文)
los	Length of stay	住院时间长短			
cost	Treatment costs	治疗花费			
clotsolv	Clot-dissolving drugs	凝块消融药	1	Streptokinase	链激酶
			2	reteplase	瑞替普酶
			3	alteplase	阿替普酶
proc	Surgical treatment	手术治疗	1	PTCA	经皮冠状动脉成形术
			2	CABG	搭桥术

数据文件中包括接受溶解血栓药物治疗的心梗患者 1481 个样本的处理记录。使用一般线性模型多变量过程对住院时间(天)和治疗药物的花费进行多元方差分析。

2) 初步分析操作步骤

(1) 读取数据文件 data09-08。按**【分析→一般线性模型→多变量】**顺序单击菜单项,打开**【多变量】**主对话框。

(2) 在主对话框中定义分析变量。

① 将住院时间变量 los 和治疗花费变量 cost 移入**【因变量】**框,作为因变量。

② 将 clotsolv 和 proc(凝块消融药和手术治疗)变量作为固定因素移入**【固定因子】**栏内。

(3) 由于要使用系统默认的全模型,因此**【模型】**对话框不用打开。

(4) 单击**【对比】**按钮,打开**【多变量:对比】**对话框,如图 9-24 所示。



图 9-24 **【多变量: 对比】**对话框



- ① 选择 clotsolv (None) 作为对比变量。
- ② 在【更改对比】栏内，单击【对比】参数框内的向下箭头，打开下拉列表，选择【简单】作为比较类型，再选择【第一个】项为比较参考类，然后单击【更改】按钮。在【因子】栏内显示的对比表达式为【clotsolv(简单(第一个))】。

(5) 在【多变量：选项】对话框中选择【描述统计量】、【功效估计】，即要求估计效应大小，以及【SSCP 矩阵】、【方差齐性检验】。

表 9-41 各单元频数

		值标签	N
凝块消融药	1	链激酶	116
	2	瑞替普酶	696
	3	阿替普酶	669
手术处理	1	经皮冠状动脉成形术	907
	2	搭桥术	574

手术类型分组的描述统计量。

表 9-43 所示是多元检验的 SSCP 矩阵。多元检验表对每个模型效应显示了 4 种显著性检验。

- Pillai 迹(表中汉化为“Pillai 的跟踪”不妥)是一个正值统计量。该统计量值越大表明对模型贡献的效应越多。
- Wilks  $\lambda$ (表中的“Wilks 的 Lambda”)是一个正值统计量，其值在 0~1 之间。统计量值越小表明效应对模型贡献越多。
- Hotelling 迹(表中汉化为“Hotelling 的跟踪”不妥)是检验矩阵特征值之和。它是一个正值统计量，值越大，表明对模型贡献的效应越多。Hotelling 迹永远大于 Pillai 迹，但当检验矩阵的特征值很小时，这两个统计量接近相等，这表明效应对模型没什么贡献。
- Roy 的最大根是检验矩阵的最大特征值，它是一个正值统计量，其值越大表明贡献给模型的效应越多。Roy 的最大根永远小于或等于 Hotelling 迹。当这两个统计量相等时该效应主要与一个因变量相联系，在因变量之间存在很强的相关性，或者该效应对模型没有什么贡献。

可以看出凝块消融药物对模型贡献不大，因为它们 Pillai 迹、Hotelling 迹、Roy 的最大根的值分别为 0.026、0.027、0.027，都很小，而 Wilks 的 Lambda 的值却很大，0.974 接近 1。

以上 4 个多元统计量都转换到具有近似或确切的 F 分布的检验统计量，给出了 F 分布的假设自由度(分子)、误差自由度(分母)以及概率值。主效应 clotsolv 和 proc 的显著性值 Sig.小于 0.05，表明效应对模型有显著性的贡献。相比之下，它们的交互项对模型没有贡献。虽然 clotsolv 对模型有贡献，因为 Pillai

(6) 单击【继续】按钮，返回主对话框。单击【确定】按钮，提交系统执行。

3) 输出结果(见表 9-41~表 9-43)

表 9-41 所示为组间因素各水平组合的单元频数，可以看出单元大小不一。

表 9-42 所示为每个因变量按服用的凝块消融药和

表 9-42 描述统计量

凝块消融药 手术处理			均值	标准 偏差	N
住院时间	1 链激酶	1 经皮冠状动脉成形术	4.94	1.105	68
		2 搭桥术	7.25	1.263	48
		总计	5.90	1.633	116
	2 瑞替普酶	1 经皮冠状动脉成形术	4.81	1.072	441
		2 搭桥术	6.62	1.137	255
		总计	5.47	1.399	696
	3 阿替普酶	1 经皮冠状动脉成形术	4.68	1.048	398
		2 搭桥术	6.48	1.135	271
		总计	5.41	1.396	669
	总计	1 经皮冠状动脉成形术	4.77	1.066	907
治疗花费	1 链激酶	1 经皮冠状动脉成形术	6.60	1.163	574
		2 搭桥术	5.48	1.422	1481
		总计	5.48	1.422	1481
	2 瑞替普酶	1 经皮冠状动脉成形术	28.3838	3.27388	68
		2 搭桥术	44.7225	5.42780	48
		总计	35.1447	9.14344	116
	3 阿替普酶	1 经皮冠状动脉成形术	29.6674	3.18096	441
		2 搭桥术	44.6251	5.22506	255
		总计	35.1476	8.27021	696
	总计	1 经皮冠状动脉成形术	29.8073	3.60094	398
协方差矩阵等同性的 Box 检验 <sup>a</sup>	1 链激酶	1 经皮冠状动脉成形术	44.7432	5.63081	271
		2 搭桥术	35.8575	8.62337	669
		总计	29.6326	3.39406	907
	2 瑞替普酶	1 经皮冠状动脉成形术	44.6890	5.42789	574
		2 搭桥术	35.4681	8.50314	1481
		总计	35.4681	8.50314	1481
	3 阿替普酶	1 经皮冠状动脉成形术	29.8073	3.60094	398
		2 搭桥术	44.7432	5.63081	271
		总计	35.8575	8.62337	669
	总计	1 经皮冠状动脉成形术	29.6326	3.39406	907
	总计	2 搭桥术	44.6890	5.42789	574
	总计	总计	35.4681	8.50314	1481

协方差矩阵等同性的 Box 检验<sup>a</sup>

迹的值与 Hotelling 迹的值接近，但其贡献不是很大。更直接的方法是看偏 $\eta^2$ (表中的“偏 Eta 方”)统计量。该统计量报告每一项的实际的显著性，是根据由效应计算的变异与由效应和剩在误差里的效应之和的比值。偏 $\eta^2$  的值较大的表明较大的模型效应，最大值为 1。由于 clotsolv 的偏 $\eta^2$  非常小，无论是对住院时间长短还是对治疗花费，变量 clotsolv 的偏 $\eta^2$  都非常小，分别为 0.015 和 0.02，说明它对模型的贡献不太大。

表 9-43 多变量检验

效应		值	F	假设 df	误差 df	Sig.	偏 Eta 方
截距	Pillai 的跟踪	.975	28781.280 <sup>b</sup>	2.000	1474.000	.000	.975
	Wilks 的 Lambda	.025	28781.280 <sup>b</sup>	2.000	1474.000	.000	.975
	Hotelling 的跟踪	39.052	28781.280 <sup>b</sup>	2.000	1474.000	.000	.975
	Roy 的最大根	39.052	28781.280 <sup>b</sup>	2.000	1474.000	.000	.975
clotsolv	Pillai 的跟踪	.026	9.833	4.000	2950.000	.000	.013
	Wilks 的 Lambda	.974	9.892 <sup>b</sup>	4.000	2948.000	.000	.013
	Hotelling 的跟踪	.027	9.952	4.000	2946.000	.000	.013
	Roy 的最大根	.027	19.909 <sup>c</sup>	2.000	1475.000	.000	.026
proc	Pillai 的跟踪	.622	1212.157 <sup>b</sup>	2.000	1474.000	.000	.622
	Wilks 的 Lambda	.378	1212.157 <sup>b</sup>	2.000	1474.000	.000	.622
	Hotelling 的跟踪	1.645	1212.157 <sup>b</sup>	2.000	1474.000	.000	.622
	Roy 的最大根	1.645	1212.157 <sup>b</sup>	2.000	1474.000	.000	.622
clotsolv * proc	Pillai 的跟踪	.004	1.508	4.000	2950.000	.197	.002
	Wilks 的 Lambda	.996	1.508 <sup>b</sup>	4.000	2948.000	.197	.002
	Hotelling 的跟踪	.004	1.509	4.000	2946.000	.197	.002
	Roy 的最大根	.004	3.022 <sup>c</sup>	2.000	1475.000	.049	.004

- a. 设计：截距 + clotsolv + proc + clotsolv \* proc
- b. 精确统计量
- c. 该统计量是 F 的上限，它产生了一个关于显著性级别的下限。

表 9-44 均值比较的结果

凝块消融药 简单对比 <sup>a</sup>		因变量	
		治疗花费	住院时间 (周)
级别 2 和级别 1	对比估算值	.593	-.382
	假设值	0	0
	差分 (估计 - 假设)	.593	-.382
	标准误差	.439	.112
	Sig.	.176	.001
	差分的 95% 置信区间		
级别 3 和级别 1	对比估算值	.722	-.516
	假设值	0	0
	差分 (估计 - 假设)	.722	-.516
	标准误差	.439	.112
	Sig.	.100	.000
	差分的 95% 置信区间		

a. 参考类别 = 1

相比较而言，proc 的偏 $\eta^2$  较大，分别为 0.291 和 0.621，这正是所期望的。外科手术是患者必须接受的治疗，导致在住院的时间效应和最终花费比服用溶解血栓剂更大。多元检验也表明 clotsolv 效应是显著的 Sig.=0.000 即小于 0.01，这意味着至少有一种药的效应与其他不同。对比的结果将表明是不同的。

表 9-44 所示是均值比较的结果，显示了 3 种凝块消融药分组的住院时间长短的均值比较和治疗花费的均值比较。变量 clotsolv 药物的第一水平 streptokinase 链激酶是指定的参考类，第二组与第一组均值比较结果说明使用 streptokinase 链激酶比

使用 reteplase 瑞替普酶多住院 0.382 天，少花费 593 美元。由于对住院时间长度的显著性值为 0.001，小于 0.05，因此可以推断这个差异不是偶然的。costs 治疗花费的显著性值大于 0.10，所以，这个差异完全可能由于随机变异引起。

第二个对比比较了第三个水平与第一水平，即 alteplase 阿替普酶的效应和 strepto-kinase 链激酶的效应。患者服用阿替普酶比服用链激酶大约平均少住院 0.516 天，治疗费用平均高出 722 美元。

由于 Length of stay 住院时间的显著性值小于 0.05，可以推断该差异并非偶然。Treatment costs 的显著性值大于 0.10，所以该差异可能完全是由于随机变异引起的。

综上所述，使用 alteplase 和 reteplase 似乎减少了患者住院时间。此外，该项减少足以弥补治疗的费用。这样，alteplase 和 reteplase 的使用应该排在 streptokinase 的前面。然而在采用这个计划之前，应该证明模型的假设检验是准确无误的。

表 9-45 (a) 所示是协方差矩阵的齐性检验结果。检验的假设是因变量遵循多元正态分布，而且方差-协方差矩阵在各种效应之间形成的单元是相等的。Box 的 M 检验的零假设是因变量协方差矩阵在各组之间是相等的。Box 的 M 检验统计量被转换为具有  $df_1$  和  $df_2$  自由度的 F 统计量。这里的检验的  $p$  值小于 0.05，拒绝原假设。这种模型的结果是不可信的。Box's M 对大数据集是敏感的，意味着当有大量观测数据时，即使偏离齐性很小也可以检测出来。此外，对偏离正态假设也很敏感。

表 9-45 (b) 所示是 Levene 检验。其假设是各因素水平组合所定义的单元之间误差方差相等。对每个因变量分别进行检验。los 住院时间长度变量的  $p$  值大于 0.10，因此不足以在这个检验中拒绝零假设(不排除在更多样本或另一个检验方法时拒绝零假设)。costs 治疗花费变量的  $p$  值小于 0.05，表明对这个变量，违反了方差相等的假设。像 Box 的 M 检验一样，Levene 检验对大数据集是敏感的。由于齐性检验的结果违反了进行多元方差分析的假设，无法得出可信的结论。如何才能得出可信的结论呢？

表 9-45 协方差矩阵和误差方差的齐性检验

协方差矩阵等同性的 Box 检验 <sup>a</sup>	
Box 的 M	270.509
F	17.908
df1	15
df2	358296.484
Sig.	.000

检验零假设，即观测到的因变量的协方差矩阵在所有组中均相等。

a. 设计：截距 +  
clotsolv + proc +  
clotsolv \* proc

(a)

	F	df1	df2	Sig.
住院时间	1.507	5	1475	.185
治疗花费	10.001	5	1475	.000

检验零假设，即在所有组中因变量的误差方差均相等。

a. 设计：截距 + clotsolv + proc + clotsolv \* proc

(b)

主体间效应的检验

源	因变量	III 型平方和	df	均方	F	Sig.	偏 Eta 方
校正模型	治疗花费	79811.122 <sup>a</sup>	5	15962.224	865.665	.000	.746
	住院时间（周）	1217.307 <sup>b</sup>	5	243.461	202.406	.000	.407
截距	治疗花费	1027759.201	1	1027759.201	55737.565	.000	.974
	住院时间（周）	25234.532	1	25234.532	20979.169	.000	.934
clotsolv	治疗花费	50.127	2	25.063	1.359	.257	.002
	住院时间（周）	26.650	2	13.325	11.078	.000	.015
proc	治疗花费	44593.620	1	44593.620	2418.407	.000	.621
	住院时间（周）	727.562	1	727.562	604.872	.000	.291
clotsolv * proc	治疗花费	50.182	2	25.091	1.361	.257	.002
	住院时间（周）	6.757	2	3.379	2.809	.061	.004
误差	治疗花费	27197.902	1475	18.439			
	住院时间（周）	1774.185	1475	1.203			
总计	治疗花费	1970083.194	1481				
	住院时间（周）	47424.000	1481				
校正的总计	治疗花费	107009.024	1480				
	住院时间（周）	2991.492	1480				

a. R 方 = .746（调整 R 方 = .745）

b. R 方 = .407（调整 R 方 = .405）

(c)

5. 检查分析条件

在这里没有列出多元方差分析的结果。先进行多元方差分析分析条件的检查。

(1) 为解决问题，作出直方图，粗略查看一下因变量的正态性。

作图步骤是按【图形→旧对话框→直方图】顺序单击菜单项，打开作直方图的对话框。将变量 `los` 住院时间长短变量移到【变量】栏内。选择【显示正态曲线】，作为比较参考。在输出窗口显示的直方图如图 9-25 (a) 所示。使用同样方法作因变量 `cost` 治疗花费的直方图如图 9-25 (b) 所示。可以看出，住院时间变量近似为正态分布，而治疗花费有明显的两个峰，每个都近似正态分布。一元方差分析表明，这是由于不同手术类型之间花费差异显著造成的。

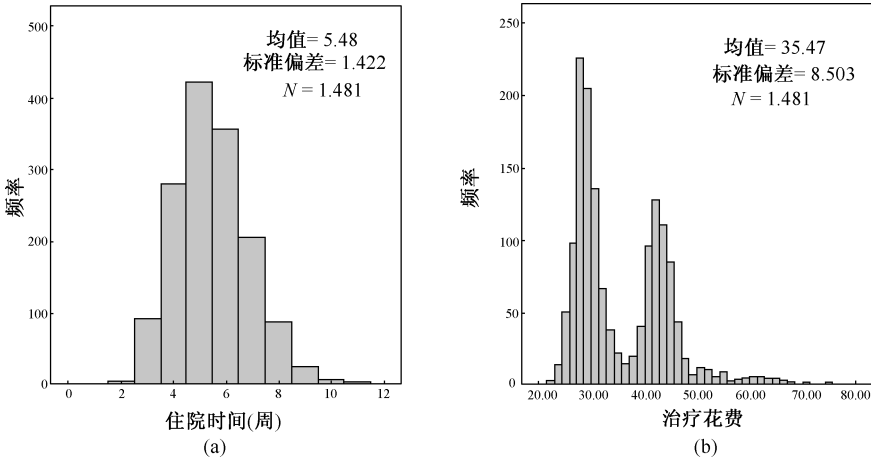


图 9-25 住院时间和治疗花费直方图



图 9-26 【分割文件】对话框

(2) 进一步分别粗略检查分类变量 `proc`，外科手术两个分类 CABG 搭桥术和 PTCA 冠状动脉成形术的花费是否是正态分布。

操作步骤是使用拆分文件功能将数据按 `proc` 的分类分开，按【数据→拆分文件】顺序打开【分割文件】对话框，如图 9-26 所示。选择【比较组】，并将变量 `proc`(手术治疗)移到【分组方式(分组依据)】栏内，单击【确定】按钮。然后按上述步骤作直方图。两类手术的花费直方图如图 9-27 所示。

这次对文件拆分的结果会一直保持到关闭该数据集为止。

(3) 虽然两种手术类型的花费都是近似正态分布的，但从图 9-27 中可以看出都左偏，因此为进一步改善正态性，对治疗花费变量 `cost` 进行常用对数转换。使用转换菜单中的计算变量命令，利用 `LG10` 函数生成新变量 `logcost = LG10(cost)`。

(4) 进行两因变量单因素的方差分析。步骤如下：

① 仍然让数据文件处于按手术类型 `proc` 分开的状态，以便并列比较。

② 按【分析→一般线性模型→多变量】顺序单击菜单项，打开主对话框，将因变量 `los` 住院天数、`logcost` 治疗花费的对数变量作为因变量移到【因变量】栏中，将 `clotsolv` 凝块消融药作为因素变量送入【固定因子】栏内。

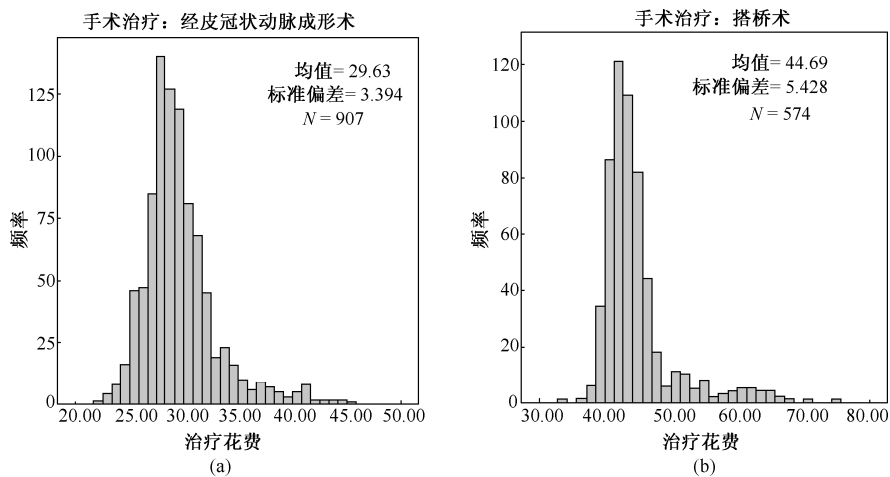


图 9-27 PTCA 和 CABG 的手术花费分布图

至此，该分析就转换成两个因变量、一个固定因子的方差分析，但是是对两类手术方法分别进行分析的。

③ 在【多变量：对比】对话框中，选择以变量 `clotsolv` 的第一水平(链激酶)作为参考类的简单比较。

④ 在【多变量：选项】对话框中，将【因子与因子交互】栏中的 (OVERALL)、`clotsolv` 送入【显示均值】栏中，并选择【比较主效应】。在【输出】栏中选择以下输出项：【描述统计量】、【功效估计】，选中【方差齐性检验】。

⑤ 在【两两比较】对话框中选择对 `clotsolv` 凝块消融药变量中各种消融药进行多重比较检验。在【假定方差齐性】栏中选择【Tukey】，在【未假定方差齐性】栏中选择【Dunnett's T3】。

(5) 主要输出结果在表 9-46～表 9-56 中。

表 9-46 组间因素各单元频数

手术治疗	凝块消融药	值标签	N
1 经皮冠状动脉成形术	1	链激酶	68
	2	瑞替普酶	441
	3	阿替普酶	398
2 搭桥术	1	链激酶	48
	2	瑞替普酶	255
	3	阿替普酶	271

表 9-47 描述统计量

手术治疗		凝块消融药	均值	标准 偏差	N
经皮冠状动脉成形术	治疗花费对数	链激酶	1.4504	.04802	68
		瑞替普酶	1.4700	.04392	441
		阿替普酶	1.4715	.04883	398
		总计	1.4692	.04671	907
	住院时间（天）	链激酶	4.94	1.105	68
		瑞替普酶	4.81	1.072	441
		阿替普酶	4.68	1.048	398
		总计	4.77	1.066	907
搭桥术	治疗花费对数	链激酶	1.6478	.04796	48
		瑞替普酶	1.6470	.04538	255
		阿替普酶	1.6478	.04910	271
		总计	1.6474	.04731	574
	住院时间（天）	链激酶	7.25	1.263	48
		瑞替普酶	6.62	1.137	255
		阿替普酶	6.48	1.135	271
		总计	6.60	1.163	574

(6) 分析与结论。

因为数据文件按手术类型分开了，产生的输出表格都是按手术类型并列的，单独得出结论。这与手术类型变量作为一个因素的结果是不同的。

表 9-46 所示为组间因素各单元频数。可以看出各单元中观测数是不相等的。

表 9-47 所示为按凝块消融药分组的因变量描述统计量。作为后面分析的参考数据。

表 9-48 所示是 Box 检验的结果。检验的零假设是：因变量的协方差矩阵在 clotsolv 不同的凝块消融的各组中相等。无论是 PTCA 经皮冠状动脉成形术，还是 CABG 搭桥术的 Sig. 值，都是计算的大于等于其  $F$  值，故认为在每组中的概率  $p$ ，表中的该值均大于 0.05，证据不足以在这个检验中拒绝零假设。协方差矩阵具有齐性。

以上两个检验结果对因素水平组合形成的单元中观测数不相等的情况非常重要。

表 9-49 所示是 Levene 检验的结果。检验的零假设是因变量在 clotsolv 不同的凝块消融药各组中的误差方差相等，是对两个因变量分别进行的检验。从表中的 Sig. 值可以看出，两种手术的花费对数和住院天数的检验结果都不足以在这个检验中拒绝零假设，故认为误差方差具有齐性。

表 9-48 协方差矩阵相等的 Box 检验

1 经皮冠状动脉成形术	Box 的 M	12.208
	F	2.021
	df1	6
	df2	246014.589
	Sig.	.059
2 搭桥术	Box 的 M	7.804
	F	1.288
	df1	6
	df2	126223.460
	Sig.	.259

检验零假设，即观测到的因变量的协方差矩阵在所有组中均相等。

a. 设计: 截距 + clotsolv

表 9-49 误差方差相等的 Levene 检验

手术治疗		F	df1	df2	Sig.
经皮冠状动脉成形术	治疗花费对数	1.950	2	904	.143
	住院时间(天)	.545	2	904	.580
搭桥术	治疗花费对数	.820	2	571	.441
	住院时间(天)	.524	2	571	.592

检验零假设，即在所有组中因变量的误差方差均相等。

a. 设计: 截距 + clotsolv

表 9-50 所示是多元检验的结果，4 种检验的统计量为 Pillai 迹、Hotelling 迹、Roy 的最大根，统计量越大，对模型贡献越大，但是表中值都很小；Wilks 的 Lambda 统计量的值越小，对模型贡献越大，而表中相应的值却很大，接近 1。所以 4 个统计量都说明因素变量凝块消融药 clotsolv 效应对模型的贡献不大。但是表中的  $F$  检验的 Sig. 值，即大于等于  $F$  值的概率都小于 0.01，而偏  $\eta^2$  值也都很小，又说明它们是有贡献的，但贡献不大。

表 9-50 对效应的 4 种检验

手术治疗			效应	值	F	假设 df	误差 df	Sig.	偏 Eta 方
1 经皮冠状动脉成形术	截距	Pillai 的跟踪		.998	294260.686 <sup>b</sup>	2.000	903.000	.000	.998
		Wilks 的 Lambda		.002	294260.686 <sup>b</sup>	2.000	903.000	.000	.998
		Hotelling 的跟踪		651.740	294260.686 <sup>b</sup>	2.000	903.000	.000	.998
		Roy 的最大根		651.740	294260.686 <sup>b</sup>	2.000	903.000	.000	.998
	clotsolv	Pillai 的跟踪		.038	8.824	4.000	1808.000	.000	.019
		Wilks 的 Lambda		.962	8.889 <sup>b</sup>	4.000	1806.000	.000	.019
		Hotelling 的跟踪		.040	8.955	4.000	1804.000	.000	.019
		Roy 的最大根		.038	17.317 <sup>c</sup>	2.000	904.000	.000	.037
2 搭桥术	截距	Pillai 的跟踪		.999	203479.448 <sup>b</sup>	2.000	570.000	.000	.999
		Wilks 的 Lambda		.001	203479.448 <sup>b</sup>	2.000	570.000	.000	.999
		Hotelling 的跟踪		713.963	203479.448 <sup>b</sup>	2.000	570.000	.000	.999
		Roy 的最大根		713.963	203479.448 <sup>b</sup>	2.000	570.000	.000	.999
	clotsolv	Pillai 的跟踪		.038	5.484	4.000	1142.000	.000	.019
		Wilks 的 Lambda		.962	5.528 <sup>b</sup>	4.000	1140.000	.000	.019
		Hotelling 的跟踪		.039	5.571	4.000	1138.000	.000	.019
		Roy 的最大根		.039	11.165 <sup>c</sup>	2.000	571.000	.000	.038

a. 设计: 截距 + clotsolv

b. 精确统计量

c. 该统计量是  $F$  的上限，它产生了一个关于显著性级别的下限。

表 9-51 所示是多元方差分析表。对住院时间(天)变量检验的零假设是：不同的凝块消融药组的平均住院时间(天数)之间无显著差异。

表 9-51 多元方差分析检验结果

主体间效应的检验								
源	因变量	III 型平方和	df	均方	F	Sig.	偏 Eta 方	
1 经皮冠状动脉成形术	校正模型	住院时间(周)	5.725 <sup>a</sup>	2	2.863	2.529	.080	.006
		治疗花费对数	.026 <sup>b</sup>	2	.013	6.112	.002	.013
	截距	住院时间(周)	10695.320	1	10695.320	9448.846	.000	.913
		治疗花费对数	989.845	1	989.845	458854.396	.000	.998
	clotsolv	住院时间(周)	5.725	2	2.863	2.529	.080	.006
		治疗花费对数	.026	2	.013	6.112	.002	.013
	误差	住院时间(周)	1023.254	904	1.132			
		治疗花费对数	1.950	904	.002			
	总计	住院时间(周)	21624.000	907				
		治疗花费对数	1959.676	907				
	校正的总计	住院时间(周)	1028.979	906				
		治疗花费对数	1.976	906				
2 搭桥术	校正模型	住院时间(周)	24.504 <sup>c</sup>	2	12.252	9.316	.000	.032
		治疗花费对数	7.466E-005 <sup>d</sup>	2	3.733E-005	.017	.984	.000
	截距	住院时间(周)	14546.869	1	14546.869	11061.280	.000	.951
		治疗花费对数	858.818	1	858.818	382466.399	.000	.999
	clotsolv	住院时间(周)	24.504	2	12.252	9.316	.000	.032
		治疗花费对数	7.466E-005	2	3.733E-005	.017	.984	.000
	误差	住院时间(周)	750.931	571	1.315			
		治疗花费对数	1.282	571	.002			
	总计	住院时间(周)	25800.000	574				
		治疗花费对数	1559.156	574				
	校正的总计	住院时间(周)	775.436	573				
		治疗花费对数	1.282	573				

a. R 方 = .006 (调整 R 方 = .003)  
b. R 方 = .013 (调整 R 方 = .011)  
c. R 方 = .032 (调整 R 方 = .028)  
d. R 方 = .000 (调整 R 方 = -.003)

对治疗花费对数变量检验的假设是：不同的凝块消融药组的平均治疗花费(以 10 为底的对数)之间无显著差异。

先看外科手术是冠状动脉成形术(PTCA)这一组，分类变量 clotsolv 的 F 检验的显著性概率，对住院时间(天)Sig. = 0.08，大于 0.05，不足以拒绝原假设。说明不同的凝块消融药组的平均住院时间之间无显著性差异。对治疗花费对数 Sig. = 0.02，小于 0.05，拒绝原假设。说明在本例条件下，不同的凝块消融药组的治疗花费均值差异显著。

再看外科手术是搭桥术(CABG)这一组，分类变量 clotsolv 的 F 检验的显著性概率，对住院时间(天)Sig. = 0.00，小于 0.05，拒绝原假设说明不同的凝块消融药组的平均住院时间之间有显著性差异。表中对治疗花费对数 Sig. = 0.982，即  $p > 0.05$ ，不足以拒绝原假设。说明在本例条件下，不同的凝块消融药组的治疗花费均值没有显著差异。

表 9-52 所示是做经皮冠状动脉成形术一组不同凝块消融药对住院时间和治疗花费影响的对比表。

第二水平与第一水平比较即使用新药瑞替普酶与使用链激酶相比，平均住院时间少了 0.129 天，Sig. = 0.351 说明这个差异是由随机因素引起的，具有一定的偶然性。平均治疗花费的对数高出 0.020，表中 Sig. = 0.001，即  $p < 0.05$ ，说明不是偶然的。转换后为高出  $10^{0.02} = 1.047$  千美元。

第三水平与第一水平比较即使用阿替普酶与使用链激酶相比，平均住院时间少了 0.258 天，Sig. = 0.065 说明这个差异是由随机因素引起的，具有一定的偶然性。平均治疗花费的对数高出 0.021，输出表中 Sig. = 0.001，即  $p < 0.05$ ，说明不是偶然的。转换后为高出  $10^{0.021} = 1.05$  千美元。

表 9-52 不同凝块消融药对住院时间和治疗花费的影响(PTCA)  
手术治疗=经皮冠状动脉成型术

凝块消融药 简单对比 <sup>a</sup>		因变量	
		治疗花费对数	住院时间 (天)
级别 2 和级别 1	对比估算值	.020	-.129
	假设值	0	0
	差分 (估计 - 假设)	.020	-.129
	标准 误差	.006	.139
	Sig.	.001	.351
	差分的 95% 置信区间		
	下限	.008	-.401
	上限	.031	.143
级别 3 和级别 1	对比估算值	.021	-.258
	假设值	0	0
	差分 (估计 - 假设)	.021	-.258
	标准 误差	.006	.140
	Sig.	.001	.065
	差分的 95% 置信区间		
	下限	.009	-.532
	上限	.033	.016

a. 参考类别 = 1

表 9-53 所示是做搭桥术一组不同凝块消融药对住院时间和治疗花费影响的对比表。

表 9-53 不同凝块消融药对住院时间和治疗花费的影响(CABG)  
手术治疗=搭桥术

凝块消融药 简单对比 <sup>a</sup>		因变量	
		治疗花费对数	住院时间 (天)
级别 2 和级别 1	对比估算值	-.001	-.634
	假设值	0	0
	差分 (估计 - 假设)	-.001	-.634
	标准 误差	.007	.180
	Sig.	.922	.000
	差分的 95% 置信区间		
	下限	-.015	-.989
	上限	.014	-.280
级别 3 和级别 1	对比估算值	-6.203E-006	-.774
	假设值	0	0
	差分 (估计 - 假设)	-6.203E-006	-.774
	标准 误差	.007	.180
	Sig.	.999	.000
	差分的 95% 置信区间		
	下限	-.015	-1.127
	上限	.015	-.421

a. 参考类别 = 1

第二水平与第一水平比较即使用新药瑞替普酶与使用链激酶相比，平均住院时间少了 0.634 天，Sig.小于 0.01，说明这个差异不是随机因素引起的。根据描述统计量表中的总平均住院天数是 6.6 天，使用新药使住院时间减少了近 10%。平均治疗花费的对数高出 0.001，Sig. = 0.922，即  $p > 0.05$ ，说明是偶然的。两种药的平均治疗花费是相等的。

第三水平与第一水平比较即使用阿替普酶与使用链激酶相比，平均住院时间少了 0.774 天，Sig.小于 0.01 说明这个差异不是由随机因素引起的。使用新药阿替普酶使住院时间减少了 11.3%。平均治疗花费的对数高出 0，表中 Sig. = 0.999，即  $p > 0.05$ ，现有证据不足以拒绝两种药的 平均治疗花费是相等的假设。

表 9-54 所示是均值多重比较结果，带有 “\*” 标记的是差异显著的两个水平的均值。从



表 9-48 和表 9-49 得出两个方差均具有齐性的结论。在观察多重比较表时应选择方差齐性的方法分析结果。在表 9-54 中就是查看“Tukey HSD”一栏的数据，进行分析。这个表比较大，观察一致性子集表更容易些。一致性子集是将多重比较综合得出的表格，更容易得出结论。

表 9-54 均值多重比较表

								95% 置信区间	
手术治疗	因变量		(I) 链块酒融酶	(J) 链块酒融酶	均值差值 (I-J)	标准 误差	Sig.	下限	上限
经皮冠状动脉成形术	治疗花费对数	Tukey HSD	链激酶	瑞替普酶	-.0196	.00605	.004	-.0338	-.0054
				阿替普酶	-.0211*	.00609	.002	-.0354	-.0068
			瑞替普酶	链激酶	.0196	.00605	.004	.0054	.0338
				阿替普酶	-.0015	.00321	.890	-.0090	.0061
			阿替普酶	链激酶	.0211*	.00609	.002	.0068	.0354
				瑞替普酶	.0015	.00321	.890	-.0061	.0090
		Dunnett T3	链激酶	瑞替普酶	-.0196	.00619	.006	-.0347	-.0045
				阿替普酶	-.0211*	.00632	.004	-.0364	-.0057
			瑞替普酶	链激酶	.0196	.00619	.006	.0045	.0347
				阿替普酶	-.0015	.00322	.956	-.0092	.0062
			阿替普酶	链激酶	.0211*	.00632	.004	.0057	.0364
				瑞替普酶	.0015	.00322	.956	-.0062	.0092
	住院时间（天）	Tukey HSD	链激酶	瑞替普酶	.13	.139	.619	-.20	.45
				阿替普酶	.26	.140	.155	-.07	.59
			瑞替普酶	链激酶	-.13	.139	.619	-.45	.20
				阿替普酶	.13	.074	.189	-.04	.30
			阿替普酶	链激酶	-.26	.140	.155	-.59	.07
				瑞替普酶	-.13	.074	.189	-.30	.04
		Dunnett T3	链激酶	瑞替普酶	.13	.143	.747	-.22	.48
				阿替普酶	.26	.144	.211	-.09	.61
			瑞替普酶	链激酶	-.13	.143	.747	-.48	.22
				阿替普酶	.13	.073	.221	-.05	.30
			阿替普酶	链激酶	-.26	.144	.211	-.61	.09
				瑞替普酶	-.13	.073	.221	-.30	.05

(a)

							95% 置信区间		
手术治疗	因变量		(I) 链块酒融酶	(J) 链块酒融酶	均值差值 (I-J)	标准 误差	Sig.	下限	上限
搭桥术	治疗花费对数	Tukey HSD	链激酶	瑞替普酶	.0007	.00746	.995	-.0168	.0183
				阿替普酶	.0000	.00742	1.000	-.0174	.0174
			瑞替普酶	链激酶	-.0007	.00746	.995	-.0183	.0168
				阿替普酶	-.0007	.00413	.983	-.0104	.0090
			阿替普酶	链激酶	.0000	.00742	1.000	-.0174	.0174
				瑞替普酶	.0007	.00413	.983	-.0090	.0104
		Dunnett T3	链激酶	瑞替普酶	.0007	.00748	1.000	-.0176	.0191
				阿替普酶	.0000	.00754	1.000	-.0184	.0185
			瑞替普酶	链激酶	-.0007	.00748	1.000	-.0191	.0176
				阿替普酶	-.0007	.00412	.997	-.0106	.0091
			阿替普酶	链激酶	.0000	.00754	1.000	-.0185	.0184
				瑞替普酶	.0007	.00412	.997	-.0091	.0106
	住院时间（天）	Tukey HSD	链激酶	瑞替普酶	.63	.180	.001	.21	1.06
				阿替普酶	.77*	.180	.000	.35	1.20
			瑞替普酶	链激酶	-.63	.180	.001	-1.06	-.21
				阿替普酶	.14	.100	.344	-.10	.37
			阿替普酶	链激酶	-.77*	.180	.000	-1.20	-.35
				瑞替普酶	-.14	.100	.344	-.37	.10
		Dunnett T3	链激酶	瑞替普酶	.63	.196	.006	.15	1.11
				阿替普酶	.77*	.195	.001	.30	1.25
			瑞替普酶	链激酶	-.63	.196	.006	-1.11	-.15
				阿替普酶	.14	.099	.405	-.10	.38
			阿替普酶	链激酶	-.77*	.195	.001	-1.25	-.30
				瑞替普酶	-.14	.099	.405	-.38	.10

基于观测到的均值。  
误差项为均值方 (错误) = 1.315。  
\*. 均值差值在 .05 级别上较显著。

(b)

表 9-55 所示是住院天数的一致性子集，对经皮冠状动脉成形术(PTCA)来说，见左表，无

论使用哪种凝块消融药的住院天数均属于同一子集；对搭桥术(CAGB)来说，见右表。两种新药的住院天数属于同一子集，链激酶属于单一子集，平均住院天数高于两组使用新药的。

表 9-56 所示是治疗花费(对数)的一致性子集，对经皮冠状动脉成形术(PTCA)来说，两种新药的治疗花费属于同一子集，链激酶属于单一子集，平均治疗花费低于两组使用新药的；对搭桥术来说，无论使用哪种凝块消融药的治疗花费均属于同一子集。

表 9-55 住院天数的一致性子集

手术治疗=经皮冠状动脉成形术				手术治疗=搭桥术			
凝块消融药		N	子集	凝块消融药		N	子集
			1				1 2
Tukey HSD <sup>a,b</sup>	阿替普酶	398	4.68	Tukey HSD <sup>a,b</sup>	阿替普酶	271	6.48
	瑞替普酶	441	4.81		瑞替普酶	255	6.62
	链激酶	68	4.94		链激酶	48	7.25
	Sig.		.085		Sig.		.650 1.000

已显示同类子集中的组均值。  
基于观测到的均值。  
误差项为均值方 (错误) = 1.132。

a. 使用调和均值样本大小 = 153.957。

b. Alpha = .05。

已显示同类子集中的组均值。  
基于观测到的均值。  
误差项为均值方 (错误) = 1.315。

a. 使用调和均值样本大小 = 105.467。

b. Alpha = .05。

表 9-56 治疗花费(对数)的一致性子集

手术治疗=经皮冠状动脉成形术				手术治疗=搭桥术			
凝块消融药		N	子集		N	子集	
			1	2		1	
Tukey HSD <sup>a,b</sup>	链激酶	68	1.4504		Tukey HSD <sup>a,b</sup>	瑞替普酶	255 1.6470
	瑞替普酶	441		1.4700		阿替普酶	271 1.6478
	阿替普酶	398		1.4715		链激酶	48 1.6478
	Sig.		1.000	.958		Sig.	.993

已显示同类子集中的组均值。  
基于观测到的均值。  
误差项为均值方 (错误) = .002。

a. 使用调和均值样本大小 = 153.957。

b. Alpha = .05。

已显示同类子集中的组均值。  
基于观测到的均值。  
误差项为均值方 (错误) = .002。

a. 使用调和均值样本大小 = 105.467。

b. Alpha = .05。

该研究项目提出的对于心梗患者服用新的凝块消融药，比使用链激酶是否可以减少住院天数以弥补治疗的高昂费用呢？

对于搭桥术，服用新药可以缩短住院时间 10%~11%，没有证据说明治疗花费的差性异。对于成形术，服用新药的住院时间与服用链激酶是一致的，花费要高出 1000 多美元。

结论是对于心梗患者，做搭桥手术可以使用新的凝块消融药 Alteplase 阿替普酶或 Reteplase 瑞替普酶代替原来常用的链激酶，可以缩短住院时间而不增加治疗费用；对做经皮冠状动脉成形的患者使用新药只能增加治疗费用，不能缩短住院时间。

以上结论也可以从多重均值比较表(表 9-54)和两对一致性子集表(表 9-55、表 9-56)得出。

9.5 重复测量设计的方差分析

9.5.1 重复测量方差分析概述

1. 重复测量方差分析的概念与重复测量方差分析的过程

最简单的重复测量方差分析是对试验对象的两次测量，例如，试验前、后各测量一次，分析试验前后样本均值间差异的显著性，从而推断试验所施加的处理或不同条件的效应。使用配对样本 T 检验也可以进行相关分析。本章介绍的是测量次数大于等于 3 的情况下的方差分析方法。

一般线性模型中的重复测量属于高级分析过程，是对同一因变量进行重复测量，可以是同一条件下进行的重复测量，目的在于研究各种处理之间是否存在显著性差异的同时，研究被试者之间的差异；也可以是不同条件下的重复测量，目的在于研究各种处理间是否存在显著性差异的同时，研究形成重复测量条件间的差异以及这些条件与处理间的交互效应。

例如，在对某种动物不同种系的繁殖试验中，使用两种种系的动物每种系若干只，在不同温度下，测量其体重、胎儿重、脂肪厚度等，可以分析种系之间(组间因素)有关繁殖指标的差异，研究不同温度(组内因素)下繁殖指标间的差异以及随温度变化各指标的变化趋势。

一般线性模型重复测量过程是对每个观测对象在不同条件下进行几次相同的测量的方差分析。使用这个一般线性模型过程，可以检验组间因素(处理)的效应和组内(重复测量)因素的效应的零假设，可以检验处理因素的效应以及重复测量因素间的交互效应。另外，还包括协变量效应，也包括组间因素的与协变量之间的交互效应。

## 2. 几个术语

- 组间因素(Between-Subjects Factor)。即处理因素，组间因素的水平把观测划分成几个组。这里的组间因素的水平是指处理的不同水平。重复测量过程研究不同水平间因变量之间差异。
- 组内因素(Within-Subjects Factor)。组内因素形成重复测量条件。组内因素的不同水平决定了对观测对象的重复测量次数。重复测量过程研究重复测量的各组间的差异。
- 测试指标名称(Measure name)。即模型中的因变量，是对每次测量的量给一个名字。
- 协变量(Covariates)。尺度类型的预测因子如果在因素水平的组合(单元)中，与因变量的值是线性相关的，应该选做模型中的协变量。
- 交互效应(Interactions)。一般线性模型重复测量过程默认产生具有全部因素交互效应的模型，这意味着因素水平的每个组合与因变量有不同的线性效应。另外，如果认为在协变量与因变量之间的线性关系因因素水平的不同而不同时，可以指定因素变量与协变量的交互效应。

例如，在一项减肥研究中，每周测量几个人的体重，共测量5周。在数据文件中，每个人是一个观测对象，或称事件。各周所测量的体重记录在变量 `weight1~weight5` 中，每个人的性别记录在另一个变量中。对每个观测对象重复测量的体重可以通过定义组内因素组织起来。该因素可叫做 `week`，定义它有5个水平。在主对话框中，变量 `weight1`， $\cdots$ ，`weight5` 被分派成 `week` 的5个水平。数据文件中，性别变量可以定义为组间变量，以便研究男女之间的差异。`weight` 作为 Measures 变量，该测量在数据文件中并不作为变量存在，但是在这里定义。有时具有多个测量的模型叫做双重重复测量模型。

## 3. 偏差平方和的分解

在重复测量设计的方差分析中的偏差平方和分解如下。以  $m$  水平的处理因素把样本观测分为  $m$  组， $j = 1 \sim m$ ；每组有  $n$  个试验对象， $i = 1 \sim n$ ；对每个试验对象进行  $l$  次测量， $k = 1 \sim l$  的重复测量试验为例。

(1) 总处理的偏差平方和被分解为处理间的偏差平方和、重复测量间的偏差平方和与处理因素与重复测量因素之间的交互的偏差平方和，有

$$S_{\text{总处理}} = n \sum_{j=1}^m \sum_{k=1}^l (\bar{x}_{jk} - \bar{\bar{x}})^2 \quad \text{自由度为 } m \cdot l - 1$$

式中,  $\bar{x}_{jk} = \frac{1}{n} \sum_{i=1}^n x_{ijk}$ ,  $\bar{\bar{x}} = \frac{1}{mnl} \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^l x_{ijk}$ 。

$$S_{\text{重复测量间}} = nm \sum_{k=1}^l (\bar{x}_k - \bar{\bar{x}})^2 \quad \text{自由度为 } l - 1$$

式中,  $\bar{x}_k = \frac{1}{l} \sum_{i=1}^n \sum_{j=1}^m x_{ijk}$ 。

$$S_{\text{处理组间}} = nl \sum_{j=1}^m (\bar{x}_j - \bar{\bar{x}})^2 \quad \text{自由度为 } m - 1$$

式中,  $\bar{x}_j = \frac{1}{n \times l} \sum_{i=1}^n \sum_{k=1}^l x_{ijk}$ 。

$$S_{\text{处理因素} \times \text{重复测量因素}} = S_{\text{总处理}} - S_{\text{处理组间}} - S_{\text{重复测量组间}} \quad \text{交互项的自由度为 } (m-1) \cdot (l-1)$$

(2) 整个样本的总的偏差平方和分解为体现个体间变异的, 各次重复测量合计的偏差平方和与体现重复测量间差异的重复测量组内偏差平方和。公式如下:

$$S_{\text{总}} = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^l (x_{ijk} - \bar{\bar{x}})^2 \quad \text{自由度为 } n \cdot m \cdot l - 1$$

$$S_{\text{组间合计}} = \frac{1}{k} \sum_{i=1}^n \sum_{j=1}^m (x_{ij} - \bar{x}_{ij})^2 \quad \text{自由度为 } m \cdot n - 1$$

式中,  $\bar{x}_{ij} = \frac{1}{n \times m} \sum_{i=1}^n \sum_{j=1}^m x_{ij}$ ; 重复测量的合计  $x_{ij} = \sum_{k=1}^l x_{ijk}$ 。

$$S_{\text{组内合计 (重复测量间)}} = S_{\text{总处理}} - S_{\text{组间合计 (观测对象间)}} \quad \text{自由度为 } m \cdot (n-1)$$

(3) 上述组间合计的偏差平方和分解为处理组间的偏差平方和与组间误差的偏差平方和。计算公式如下:

$$S_{\text{组间误差}} = S_{\text{组间合计}} - S_{\text{处理组间}} \quad \text{自由度为 } m \cdot (n-1)$$

(4) 上述组内合计的偏差平方和可分解为重复测量间的偏差平方和、交互作用的偏差平方和与组内误差的偏差平方和。前三项根据前面公式可以计算出, 故有:

$$S_{\text{组内误差}} = S_{\text{组内合计}} - S_{\text{处理因素} \times \text{重复测量因素}} \quad \text{自由度为 } m \cdot (n-1) \cdot (l-1)$$

(3)、(4) 中的偏差平方和除以各自的自由度可得到相应的均方。它们与误差均方之商即为 F 检验的 F 值。

#### 4. SPSS 中重复测量方差分析的假设检验

假设有  $k$  个样本, 即是对同一组观测对象在  $k$  个条件下的重复测量。

(1) 原假设  $H_0$ :  $k$  次重复测量的样本均数都相同即  $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k = \mu$ ,  $k$  个样本有共同的方差  $\sigma$ , 则  $k$  个样本来自具有共同的方差  $\sigma$  和相同的均数  $\mu$  的总体。SPSS 将  $k$  次重复测量样本视为  $k$  个因变量, 做 4 种多元检验。如果经过检验  $F$  值远远大于临界值,  $p < 0.05$ , 推翻原假设,  $p > 0.05$  无法拒绝原假设, 样本来自相同总体, 处理无作用, 即  $k$  次重复测量之间无显著差异。

(2) 如果定义了组间因素变量, 在重复测量方差分析中, 组间偏差平方和即反映了该分组变量各水平间的差异。检验的零假设是  $H_0$ : 该分类变量各水平组成的样本来自均值相同的总体。如果经过计算, 组间均方远远大于误差均方,  $F$  值远远大于临界值,  $F_b > F_{0.05df_b, df_{sc}}$ , 则  $p < 0.05$ , 推翻原假设, 说明分组变量各水平的因变量均值差异显著, 样本来自不同的正态总体, 否则无法拒绝原假设。不足以在这个检验中拒绝零假设(不排除在更多样本时, 或另一个检验方法时拒绝零假设)。

5. 趋势分析

如果重复测量的条件是有序变化的, 例如是在不同时间点或不同温度点下进行的测量, 可以分析因变量均值随时间变化的趋势或随温度变化的趋势是线性的、二次的、三次的或更高次的。

9.5.2 重复测量方差分析的数据文件结构

1. 重复测量设计的数据及数据文件结构

在试验中进行重复测量的因变量应该是等间隔测度的(连续的)数值型数据。这些重复测量的因变量可以是在不同条件下对同一组观测对象进行的测量结果, 组合后作为组内因素, 这是重复测量设计所必需的。

在试验中的分类变量体现了观测对象的分组, 不同组观测对象在方差分析中作为组间因素。最简单的方差分析中可以不包括组间因素。

要进行重复测量方差分析, 数据的组织与其他类型的方差分析有所不同, 它要求对被试者的若干次重复测试结果作为不同因变量出现在数据文件中。例如, 教育心理研究中的对刺激反应时测量的试验方法的研究中, 设置了 3 个级别的视觉刺激作为处理因素变量, 4 位被试者均接受 3 个级别的刺激, 每个被试者给予一个编号, 该变量不参与分析, 只为输入数据及核对时使用。对每个被试者在同样条件下测量 3 次, 原始试验数据记录见表 9-57, 数据文件为 data09-09。

在数据窗口中建立数据文件的变量说明: number 被试者编号, vsno 视觉刺激等级(1=刺激 1, 2=刺激 2, 3=刺激 3), time1 反应时测量 1, time2 反应时测量 2, time3 反应时测量 3。数据文件中的数据样例如图 9-28 所示, 数据文件的结构对重复测量方差分析很重要, 一定要把每次测量值作为一个变量, 否则无法使用 SPSS 的重复测量方差分析功能对数据进行分析。

表 9-57 重复测量的原始数据

受试者	刺激 1				刺激 2				刺激 3			
	1	2	3	4	5	6	7	8	9	10	11	12
反应时测量 1	0.9	1.5	0.5	0.8	2.4	1.9	2.9	2.4	1.5	2.1	1.1	1.6
反应时测量 2	1.2	1.1	0.8	1.3	2.8	2.4	3.3	2.8	1.2	1.9	1.5	1.8
反应时测量 3	0.7	0.8	0.5	0.9	2.1	2.2	2.7	2.9	1.9	2.2	1.0	1.3

	number	vsno	time1	time2	time3
1	1	1	.9	1.2	.7
2	2	1	1.5	1.1	.8
3	3	1	.5	.8	.5
4	4	1	.8	1.3	.9
5	1	2	2.4	2.8	2.1
6	2	2	1.9	2.4	2.2
7	3	2	2.9	3.3	2.7
8	4	2	2.4	2.8	2.9
9	1	3	1.5	1.2	1.9
10	2	3	2.1	1.9	2.2
11	3	3	1.1	1.5	1.0
12	4	3	1.6	1.8	1.3

图 9-28 重复测量数据文件结构

## 2. 重复测量方差分析的假设条件

重复测量设计中的每一元每组测量中的观测应该是独立的,并符合多元正态分布,这与一元(ANOVA)、多元(MANOVA)分析一样。如果违反多元正态分布或观测独立的假设,可能得到不可解释的结果。

重复测量设计还要求满足球形假设,即每组之间的方差-协方差矩阵相等,但在各组观测相等的情况下,对该假设条件的要求并不严格。

### 9.5.3 组内因素的设置与重复测量方差分析过程

重复测量方差分析的功能模块调用步骤如图 9-2 所示,即按【分析→一般线性模型→重复度量】顺序单击菜单项,打开【重复度量定义因子】对话框,如图 9-29 所示。

#### 1) 组内因素的定义

**注意:** 这里定义的不是数据文件中的变量,而是重复测量的变量组的代号。

(1) 定义组内因素。当在菜单中选择了【重复度量】时,并不马上打开重复测量方差分析的主对话框,而是先显示定义组内因素的对话框,如图 9-29 所示。

下面研究的组内因素是由 3 个视觉刺激反应时构成的,这个组内因素命名为 time,共有 3 个水平。

① 在【重复度量定义因子】对话框的【被试内因子名称】框中输入组内因素(被试内因子)名 time,代替原显示的因子 1。

② 在【级别数】(应为【水平数】)框中输入因素水平数 3,如图 9-29 所示。输入结束后,【添加】按钮加亮,单击该按钮,定义表达式 time(3)显示在大矩形框中。如果研究的课题中还有另外的组内因素,可以用同样方法继续定义。

③ 已经定义的组内因素(被试内因子)若有错误,单击出错的定义表达式,此时【更改】、【删除】按钮加亮。单击【更改】按钮,表达式分解为组内因素名和水平,两部分分别显示在【被试内因子名称】和【级别数】框中。在框中修改后,单击【更改】按钮,正确的表达式将显示在大矩形框中。删除已经定义并显示在大矩形框中的组内因素,可以单击该表达式,然后单击【删除】按钮。

④ 如果对每个组内因素(被试内因子)所代表的变量的测量仍有重复,在【重复度量定义因子】对话框定义因子的下半部分,定义表示重复测试的变量,见图 9-29。

例如,在减肥的研究中,20 个人为试验对象。在 5 周中,每周测一次体重。在数据文件中每个人是 1 个观测(1 个记录,占 1 行)。每周测得的体重数为 5 个变量 weight1~weight5 的值。另外用 gender 变量记录他(她)们的性别。对每个人来说,体重就是重复测量的。把 weight1~weight5 定义为被试内因子名,可以命名为 week,它有 5 个水平。在【被试内因子名称】栏中输入“week”,在【级别数】栏输入“5”,单击【添加】按钮。在主对话框中,weight1~weight5 用于给 week 的 5 个水平赋值。性别变量可以定义为组间因素,以便研究男、女在减肥试验中的差异。如果还要每天测一次脉搏和呼吸,则应该在【度量名称】栏内定义这些测量结果。

(2) 进入重复测量方差分析的主对话框。检查所有定义的被试内因子表达式,正确无误后,单击【定义】按钮,结束组内因素定义工作,进入【重复度量】主对话框,如图 9-30 所示。

主对话框中有 4 个矩形框:

① 左边的矩形框显示了在数据文件中输入的所有变量。

② 右边的【群体内部变量】框下显示了在【重复度量定义因子】对话框中定义的所有被试内因子的名称。在其下的矩形框内显示待定的因素水平。

③ 【因子列表】框，是组间因素框。

④ 【协变量】框等待输入协变量。



图 9-29 【重复度量定义因子】对话框



图 9-30 【重复度量】主对话框

## 2) 定义组内因素(被试内因子)各水平组合与原始变量的对应关系

在【群体内部变量】框中显示有一系列“\_?(n)”，表示组内变量第  $n$  个水平。

在原始变量表中选择读者认为是组内因素第  $n$  水平的变量并单击。本例选择原始变量 `time1` 作为组内因素 `time` 的第一水平，因此单击左边矩形框中的“`time1`”，然后单击向右箭头按钮，右边矩形框中的“\_?(1)”变为“`time1(1)`”。如果想要让 `time1` 作 `time` 变量的第二因素，可以单击向下箭头按钮，使“`time1(1)`”变为“`time1(2)`”，即可以使用上、下箭头按钮改变组内因素变量水平与原始变量的对应关系。

**注意：**组内因素(被试内因子)水平组合表达式的括号内是水平组合。本例只定义了 1 个组内因素 `time`，因此表达式括号内为 1 个数字；如果定义了 2 个组内因素，表达式括号内为 2 个用逗号隔开的水平序号。

例如，如果定义了组内因素  $x$ ，有 2 个水平； $y$  有 3 个水平。在组内因素矩形框中将出现以下要求定义的表达式：\_?(1,1)、\_?(1,2)、\_?(1,3)、\_?(2,1)、\_?(2,2)、\_?(2,3)。

当然，在数据文件中与之对应的应该有 6 个因变量，每个因变量对应着一种水平组合。读者不难反过来思考这种重复测量方差分析设计。

## 3) 定义方差分析的组间因素变量

在变量列表中选择组间因素变量，如选择 `Vsno`，送入【因子列表】框中。

## 4) 定义协变量及其类型

如果有协变量，则在左边的变量框选中协变量，单击向右箭头按钮，送入【协变量】框中。

## 5) 定义分析模型

在主对话框中，单击【模型】按钮，打开【重复度量：模型】对话框，见图 9-31。

(1) 在指定模型栏中选择定义模型的方式。

① 【全因子】。即饱和模型，是系统默认方式。【全因子】模型包含所有因子主效应、所有协变量主效应以及所有因子间交互效应，但不包含协变量交互项。



图 9-31 【重复度量：模型】对话框

菜单【转换】中的【计算变量】功能使两个协变量相乘建立新变量，再在此建立指定新变量的各种效应。

(3) 选择计算组间模型平方和方法。在【平方和】框内可以选择分解平方和的方法。第Ⅲ类方法是常用的也是系统默认的方法。

6) 其他功能

有关【对比】功能、【绘制】功能、【两两比较】功能、【保存】功能、【选项】功能选项，均与单因变量多因素方差分析的选项、含义相同。

只有【选项】对话框中的【SSCP 矩阵】项与单因变量多因素方差分析不同。

【SSCP 矩阵】项对设计中的每个效应给出平方和与叉积矩阵。单因变量多因素方差分析中，对所有组间效应只给出一个误差阵。只有对重复测量，对每个组间效应既给出假设的 SSCP 矩阵，也给出误差 SSCP 矩阵。

7) 运行

各子对话框中的选项选定后，在各子对话框中单击【继续】按钮，返回主对话框，单击【确定】按钮，提交系统执行。

9.5.4 重复测量方差分析实例

【例 10】 下面以一元重复测量分析为例说明重复测量设计方差分析原理。研究 4 种药物对某生化指标的作用，5 名被试者参与试验，数据见表 9-58。每种药物试验之间的相隔时间足以避免药物相互之间的影响。这是无处理(条件)分组，一个组内因素的试验设计。数据文件为 data09-10。

研究的零假设为，4 种药物对某生化指标作用(组内)无显著性差异。

1) 操作步骤

(1) 按【分析→一般线性模型→重复度量】顺序单击菜单项，打开【重复度量定义因子】对话框。

②【设定】。自定义方式。选择【设定】可以仅指定一部分分析中感兴趣的交互项或指定因子与协变量的交互。选择自定义方式激活 4 个框，可以定义组内模型。

(2) 自定义模型。在【主体内】框中列出了组内因素变量。【群体间】框中列出了组间因素变量。对应的右边两个矩形框分别是【群体内(被试内)模型】和【群体间(被试间)模型】。中间的【构建项】栏是对应效应类型的下拉列表，其中有主效应、各级交互效应。选择一种类型使用，参见 9.3.2 小节中叙述的方法定义被试内模型和被试间模型。注意，如果有两个以上协变量，不能指定协变量与协变量之间的交互效应，但可以使用主

表 9-58 不同药物对受试者的影响

受试者	药物 1	药物 2	药物 3	药物 4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30



(2) 定义组内因素

① 在【重复度量定义因子】对话框的【被试内因子名称】框中删除原有的因子1，输入被试内(组内因素)因子名称“med”。注意，这里的“med”不是数据文件中的变量名，是变量med1、med2、med3、med4这组重复测量的变量的代号。

② 在【级别数】框中输入组内因素(被试内因子)med的水平数“4”。

③ 单击【添加】按钮，在大矩形框中显示【med(4)】，【定义】按钮变亮，组内因素设置完成。

④ 单击【定义】按钮，确认以上一个组内因素变量的单元，打开【重复度量】主对话框。

(3) 在主对话框中设置分析变量。设置组内因素 med 与原始变量之间的对应关系：在左边矩形框中选择 med1~med4 这 4 个变量，单击向右箭头按钮，在【群体内部变量】栏中，第一项变为【med1(1)】，第二个组内因素 med2(2)和第三、四个组内因素 med3(3)和 med4(4)。

(4) 本例的重复测量试验设计没有考虑协变量，故无须对【协变量】框进行操作。

(5) 根据数据特点，只能检验 4 种药物对生化指标变化的差异的显著性，以及受试者间的差异性，无须指定分析模型。根据试验目的无须进行均值比较。也无须进行【对比】对话框的操作。

(6) 输出选项。在主对话框中，单击【选项】按钮，打开【选项】对话框。为在输出中显示观测均值，在【因子与因子交互】框中选择“med”项，单击向右箭头按钮，使 med 显示在【显示均值】框中。选择【描述统计】复选项，目的是要求输出各种药物作用下生化指标的描述统计量，以便根据输出得出结论。

(7) 在主对话框中单击【确定】按钮，提交系统执行。

2) 输出结果(见表 9-59~表 9-61)

表 9-62 所示为比较而作的单因变量方差分析结果。

3) 结果解释与分析

表 9-59(a)所示是组内因素基本数据信息。组内因素 med，有 4 个水平，作为 4 个因变量 med1、med2、med3 和 med4。

表 9-59(b)所示是重复测量变量的描述统计量，是每种药物作用后生化指标均值、标准差及观测数。

表 9-59 基本信息与描述统计量

度量:  
MEASURE\_1

med	因变量
1	med1
2	med2
3	med3
4	med4

(a)

	均值	标准 偏差	N
服药物1后生化指标	26.40	8.764	5
服药物2后生化指标	25.60	6.542	5
服药物3后生化指标	15.60	3.847	5
服药物4后生化指标	32.00	8.000	5

(b)

表 9-60 所示为多变量检验结果。4 种方法的 F 检验的概率 Sig 值均为 0.034，小于 0.05，说明 4 种药物对该生化指标的作用差异显著。常用的 Wilks 的 Lambda 应该是 0~1 之间的值，其值越接近 0，越拒绝作为因变量的 4 种药物对某生化指标作用无差异的假设。现在的值是 0.023，也说明了拒绝原假设的结论。Roy 的最大根是检验矩阵的最大特征值，其值越大表明贡献给模型的效应越多，本例中的值为 42.618。Roy 的最大根永远小于等于 Hotelling 迹。当这两

个统计量相等时说明在因变量之间存在很强的相关性，以重复测量的 4 个变量 med1~med4 作相关分析，可以证明这个结论。Hotelling 迹永远大于 Pillai 迹。这两个统计量相距越大，表明该效应对模型贡献越多。而本例中的 Pillai 迹值为 0.977，Hotelling 迹值为 42.618，也说明因变量的 4 种药物对某生化指标这个因变量的贡献是比较多的。

表 9-60 药物间多元检验

效应		值	F	假设 df	误差 df	Sig.
med	Pillai 的跟踪	.977	28.412 <sup>b</sup>	3.000	2.000	.034
	Wilks 的 Lambda	.023	28.412 <sup>b</sup>	3.000	2.000	.034
	Hotelling 的跟踪	42.618	28.412 <sup>b</sup>	3.000	2.000	.034
	Roy 的最大根	42.618	28.412 <sup>b</sup>	3.000	2.000	.034

a. 设计 : 截距  
主体内设计: med  
b. 精确统计量

表 9-61 所示为对组内效应的方差分析结果，即重复测量间的差异。第 1 行是在满足球形假设条件下，对 F 分子、分母自由度不作调整的条件下的检验结果。下面 3 行是在不满足球形假设时 3 种不同的检验方法，对 F 检验的分子、分母自由度作了不同的调整的检验结果。4 条件下的 F 检验对应的 Sig. 值，即  $p < 0.05$ ，因此拒绝组内因素无差异的原假设，说明被试者对不同药物的反应差异有统计意义上的显著性。由于平衡设计对球形假设条件没有严格要求，而且在表中几种情况的检验结果都相同，故没有给出球形假设的输出表。

表 9-61 被试者内效应方差分析结果

源		III 型平方和	df	均方	F	Sig.
med	采用的球形度	698.200	3	232.733	24.759	.000
	Greenhouse-Geisser	698.200	1.815	384.763	24.759	.001
	Huynh-Feldt	698.200	3.000	232.733	24.759	.000
	下限	698.200	1.000	698.200	24.759	.008
误差 (med)	采用的球形度	112.800	12	9.400		
	Greenhouse-Geisser	112.800	7.258	15.540		
	Huynh-Feldt	112.800	12.000	9.400		
	下限	112.800	4.000	28.200		

表 9-62 所示为按完全随机设计进行方差分析(根据数据文件 data09-10a)的结果。与按重复测量方差分析结果比较，可以看出，按重复测量方差分析的  $F$  值远远大于按完全随机设计来分析的  $F$  值，按重复测量方差分析显著性概率更是远离 0.01。

表 9-62 一元完全随机设计方差分析

服药物1后生化指标					
	平方和	df	均方	F	显著性
组间	698.200	3	232.733	4.692	.016
组内	793.600	16	49.600		
总数	1491.800	19			

可以看出，较少的试验对象反复使用，不但可以减少人力、财力在试验中的消耗，而且可以很好地减少由于试验对象个体偏差引起的误差方差。当然，需要避免的是两次试验间的相互影响。例如，本例中，对同一个试验对象给 4 种药物进行试验，两种药物试验间的时间间隔可能要相当长，以避免前一种药物对后一次试验的影响。这是专业问题，不是统计方法问题，需要读者注意。

### 9.5.5 关于趋势分析

#### 1. 趋势分析的概念

当重复测量的条件是某些顺序变量时，可以分析重复测量的因变量随顺序变量变化的趋势。

**【例 11】** 选择 16 名试验对象，使用两种方法锻炼他们的记忆。训练一段时间后，每隔一天后每天测试一次记忆情况，共测试 5 次。每次测试对每个参与试验的人员均按一定法则打分。数据文件为 data09-11。这是一个组内因素、一个组间因素的重复测量设计的例题。因为组内因素是与时间有关的变量，因此不但可以分析比较两种训练记忆的方法哪个更有效，还可以得到随时间的推移，记忆分数随时间下降的数学模型。如果回忆的下降在整个测量的时间段上是个常数，则会发现记忆的下降与时间之间是线性关系。如果回忆的下降表现在前两天，第 3 天开始则急剧下降，会得到一个二次趋势；如果在第一天表现为下降，在以后的几天急剧下降，最后达到稳定，则回忆与时间的关系呈现为三次关系，见图 9-32。

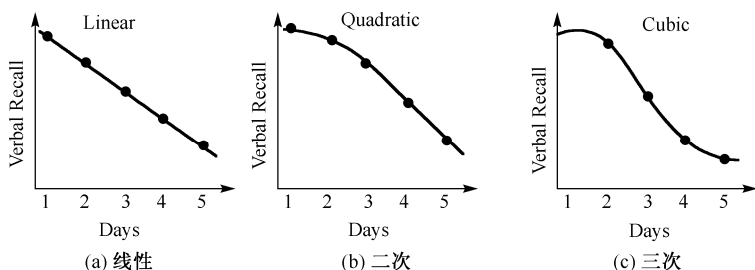


图 9-32 线性、二次、三次趋势图

#### 2. 有关记忆趋势分析的操作步骤

(1) 打开数据文件 data09-11。按【分析→一般线性模型→重复度量】顺序单击菜单项，打开【重复度量定义因子】对话框。

(2) 在【被试内因子名称】框中输入“days”，设置重复测量变量集名称为“days”。在【级别数】框中输入“5”，单击【添加】按钮，再单击【定义】按钮。

(3) 在【重复度量】主对话框中，左边栏选择 5 个因变量 day1~day5，单击向右箭头按钮，右边栏显示“day1(1), day2(2), ..., day5(5)”。将变量 group 送入【因子列表】栏中，作为被试间因子即组间变量。

(4) 单击【模型】按钮，在【模型】对话框中只选择【设定】，进行自定义模型，【构建项】栏的菜单中选择【主效应】，将【主体内】栏中的“days”作为组内因素送入【群体内模型】栏中，将【群体间】栏中的分组变量 group 送入【群体间模型】栏中。单击【继续】按钮返回主对话框。

(5) 单击【绘制】按钮，在打开的【轮廓图】对话框中，选择“days”作横轴变量送入【水平轴】栏，将 group 送入【单图】栏中作为分线变量。单击绘制窗口的【添加】按钮，确定要输出的图形表达式 days\*group。

(6) 单击【选项】按钮，在打开的【选项】对话框中，选择“days”、“group”、“(OVERALL)”送入【显示均值】栏。在输出复选项中选择【描述统计量】和【功效估计】项。

因组间变量 group 只有 2 个水平，不能进行均值的多重比较。

3. 输出结果(见表 9-63~表 9-68、图 9-33)

表 9-63 组内因素和组间因素清单

度量: MEASURE_1						
因变量						
days						
1	2	3	4	5		
day1	day2	day3	day4	day5		

(a)

		N
实验方法分组	1	8
	2	8

(b)

4. 输出结果说明

表 9-63 (a) 显示了组内因素 days 由 5 个因变量组成, 表 9-63 (b) 显示了组间因素是按试验方法分为两组, 每组 8 个试验对象。

表 9-64 描述了统计量, 每个因变量按试验组对照组分组显示均值标准差、观测数 N。

表 9-64 组合变量各水平及总描述统计量

实验方法分组		均值	标准 偏差	N
一天后分数	1	34.25	6.228	8
	2	35.00	5.928	8
	总计	34.63	5.886	16
两天后分数	1	30.88	6.728	8
	2	31.63	5.097	8
	总计	31.25	5.779	16
三天后分数	1	24.50	4.986	8
	2	24.88	4.704	8
	总计	24.69	4.686	16
四天后分数	1	19.13	5.592	8
	2	20.25	3.882	8
	总计	19.69	4.686	16
五天后分数	1	16.88	5.890	8
	2	15.25	5.651	8
	总计	16.06	5.639	16

表 9-65 所示是针对 5 天作为 5 个因变量进行的多元检验。检验的假设是 5 天里每天得分的均值相等, 各天得分与教法之间无交互作用。可以看出, 两种教学方法的 F 检验  $p$  值(表中的 Sig.)均为 0.00, 小于 0.001, 因此 5 天之间得分均值间差异显著, 从 Wilk 的 Lambda 值为 0.059, 接近 0 也可以得出同样结论。同时看出, 5 天与教法之间交互效应的 4 种检验结果  $p$  值(表中的 Sig.)都大于 0.05, 说明教法与测试延续时间之间无交互效应。

表 9-65 多元检验结果

效应		值	F	假设 df	误差 df	Sig.	偏 Eta 方
days	Pillai 的跟踪	.941	43.509 <sup>b</sup>	4.000	11.000	.000	.941
	Wilks 的 Lambda	.059	43.509 <sup>b</sup>	4.000	11.000	.000	.941
	Hotelling 的跟踪	15.821	43.509 <sup>b</sup>	4.000	11.000	.000	.941
	Roy 的最大根	15.821	43.509 <sup>b</sup>	4.000	11.000	.000	.941
days * group	Pillai 的跟踪	.364	1.573 <sup>b</sup>	4.000	11.000	.249	.364
	Wilks 的 Lambda	.636	1.573 <sup>b</sup>	4.000	11.000	.249	.364
	Hotelling 的跟踪	.572	1.573 <sup>b</sup>	4.000	11.000	.249	.364
	Roy 的最大根	.572	1.573 <sup>b</sup>	4.000	11.000	.249	.364

a. 设计: 截距 + group  
主体内设计: days

b. 精确统计量

表 9-66 所示为组内因素效应检验，无论是否符合球形假设的前提条件，4 种方法计算的组合变量 days 四类偏差平方和相等，只是根据不同方法调整了自由度，F 检验的结果出现当前 F 值及其更加极端值的概率均小于 0.001，说明 5 天之间的分数均值差异显著，days 与 group 交互效应不显著。

表 9-66 组内因素效应检验结果

度量: MEASURE\_1

源		III 型平方和	df	均方	F	Sig.	偏 Eta 方
days	采用的球形度	3832.925	4	958.231	135.268	.000	.906
	Greenhouse-Geisser	3832.925	1.870	2049.340	135.268	.000	.906
	Huynh-Feldt	3832.925	2.302	1664.885	135.268	.000	.906
	下限	3832.925	1.000	3832.925	135.268	.000	.906
days * group	采用的球形度	19.175	4	4.794	.677	.611	.046
	Greenhouse-Geisser	19.175	1.870	10.252	.677	.507	.046
	Huynh-Feldt	19.175	2.302	8.329	.677	.535	.046
	下限	19.175	1.000	19.175	.677	.425	.046
误差 (days)	采用的球形度	396.700	56	7.084			
	Greenhouse-Geisser	396.700	26.184	15.150			
	Huynh-Feldt	396.700	32.231	12.308			
	下限	396.700	14.000	28.336			

表 9-67 所示是利用对比进行的趋势分析结果。假设：①分数均值随天数变化的趋势不具有线性特性；②不具有二次特性；③不具有三次特性；④不具有四次特性。各种回归分析的方差分析表明除②外均可在这个检验中拒绝零假设，因为① $p < 0.001$ ，② $p = 0.623 > 0.05$ ，③ $p = 0.003 < 0.01$ ，④ $p = 0.039 > 0.01$ ，但 $p < 0.05$ 。因此，可以认为得分随时间变化的趋势符合线性或三次、四次函数的特征。作为结论，犯错误的概率小于 0.01 或 0.05。根据偏 $\eta^2$ 的大小，可以选择认为变化趋势为线性下降的。

表 9-67 组内因素多项式对比检验结果(趋势分析)

度量: MEASURE\_1

源	days	III 型平方和	df	均方	F	Sig.	偏 Eta 方
days	线性	3792.756	1	3792.756	193.729	.000	.933
	二次	1.290	1	1.290	.253	.623	.018
	三次	33.306	1	33.306	12.850	.003	.479
	阶 4	5.572	1	5.572	5.197	.039	.271
days * group	线性	7.656	1	7.656	.391	.542	.027
	二次	5.469	1	5.469	1.074	.318	.071
	三次	3.906	1	3.906	1.507	.240	.097
	阶 4	2.144	1	2.144	1.999	.179	.125
误差 (days)	线性	274.088	14	19.578			
	二次	71.313	14	5.094			
	三次	36.287	14	2.592			
	阶 4	15.012	14	1.072			

表 9-68 所示为边际均值表。左表是按试验方法分组的均值、标准误和 95%置信区间；右表是按时间顺序分组的均值、标准误和 95%置信区间。

图 9-33 所示是每天平均分数图。图中的每种方法对应一条折线。可以看出，试验方法 2 随时间的推移记忆的得分下降趋势近似直线，试验方法 1 的折线近似于三次曲线，这与表 9-67 中趋势分析的综合结果也是相符合的。

表 9-68 边际均值表

3. 实验方法分组

度量: MEASURE_1				
实验方法分组	均值	标准 误差	95% 置信区间	
			下限	上限
1	25.125	1.762	21.345	28.905
2	25.400	1.762	21.620	29.180

2. days

度量: MEASURE_1				
days	均值	标准 误差	95% 置信区间	
			下限	上限
1	34.625	1.520	31.365	37.885
2	31.250	1.492	28.050	34.450
3	24.688	1.212	22.089	27.286
4	19.688	1.203	17.107	22.268
5	16.063	1.443	12.968	19.157

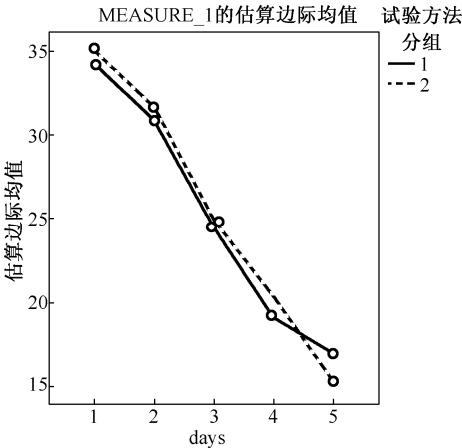


图 9-33 趋势图(边际均值图)

9.6 方差成分分析

方差成分分析是研究混合效应模型中各随机效应对因变量方差的贡献。这个过程主要适用于对混合模型的分析，如对裂区、单变量重复测量和随机区组设计的分析。通过计算方差成分，可以找出减小方差的方向。方差成分分析过程共有 4 种分析方法：最小正规二次无偏估计(MINQUE)法、方差分析(ANOVA)法、最大似然(ML)法和有限最大似然(REML)法。各种方法的默认输出项都包括方差成分估计值。如果使用最大似然法和有限最大似然法还输出渐近协方差阵。其他输出还包括方差分析表和方差分析的期望均方，使用最大似然法和有限最大似然法还输出迭代过程。

方差成分分析过程与一般线性模型(GLM)单因变量方差分析过程完全兼容。

WLS 权重允许指定一个加权变量，进行加权分析时，用于给各观测不同的权重，或作为不同测量精度的补偿。

方差成分分析要求因变量是数值变量。因素变量是分类变量，既可以是数值型变量，也可以是最多由 8 个字符组成的字符型变量。至少要有一个因素是随机的。也就是说，该因素的水平必须是从可能的水平中随机采样得来的。协变量是数值型变量，并与因变量有一定的相关关系。

所有方差成分分析方法都假设：随机效应模型参数的均值为 0 和方差为有限常数，并且彼此不相关。不同效应的模型参数也不相关。

残差项也有零均值和有限常数方差。它与任意一个随机效应的模型参数都不相关。不同观测的残差项也假设为彼此不相关。

根据这些假设,随机因素同一水平的观测是彼此相关的。ANOVA 法和 MINQUE 法不要求正态假设。虽然它们都是在正态假设条件下的方法,但它们都能缓解违反正态分布带来的影响。

ML 法和 REML 法要求模型参数和残差项服从正态分布。

在进行方差成分分析之前,可以使用探索(Explore)过程检测数据。可以使用一般线性模型的单变量、多变量和重复度量几个过程进行假设检验。

### 9.6.1 方差成分分析过程

方差成分的功能模块调用步骤见图 9-2,即按【分析→一般线性模型→方差分量估计】顺序单击菜单项,打开【方差成分分析】主对话框,见图 9-34。

#### 1. 定义因变量和随机因子

注意在作方差成分分析时一定要指定随机因子。在左边的变量框中选择因变量,单击向右箭头,将其送入【因变量】框,再从左边的变量框中选中随机因子,单击向右箭头,将其送入【随机因子】框。

如果需要协变量分析,可以指定协变量进入【协变量】框。如果需要分析权重,可以指定权重因子进入【WLS 权重】框。完成以上工作后即可以通过各功能按钮打开相应的对话框,可选择【模型】、【选项】、【保存】等按钮。



图 9-34 【方差成分】主对话框

#### 2. 【方差成分：模型】对话框

单击【模型】按钮,打开【方差成分：模型】对话框,见图 9-35。选择分析模型。如果选择【全因子】,模型中包括所有因素变量主效应、协变量主效应、因素变量之间的交互效应,但不包括协变量交互项。如果要自定义模型,选择【设定】项,可以指定因素变量与协变量的交互效应。模型中必须包括随机因素变量。

#### 3. 【方差成分：选项】对话框

单击【选项】按钮,打开【方差成分：选项】对话框,选择分析方法,见图 9-36。

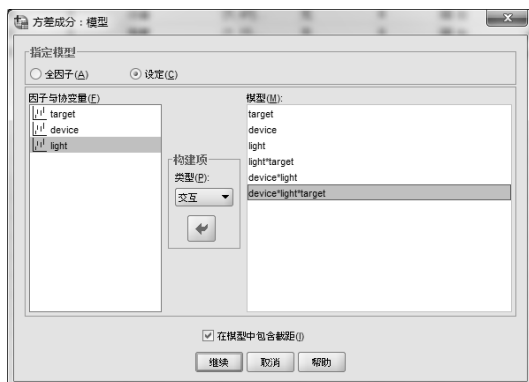


图 9-35 【方差成分：模型】对话框

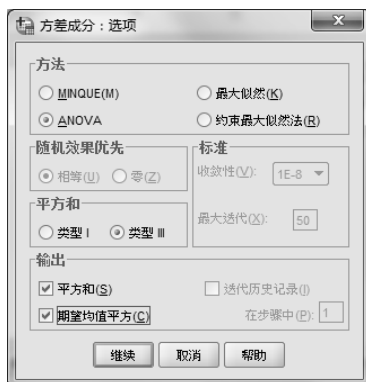


图 9-36 【方差成分：选项】对话框

(1) 在【方法】栏内指定一种进行方差成分分析的方法。有 4 种方法可供选择。

①【MINQUE】。正态最小二次无偏估计,就固定效应而言,产生的估计是不变的。如果数据是正态分布的且估计是正确的,则使用此方法作方差大小估计要比其他方法得到的方差小。这是系统默认方法。

②【ANOVA】。对每个效应使用 Type I 或 Type III 平方和分解方法进行无偏估计。ANOVA 方法有时产生负方差估计,这表明模型不正确或估计方法不合适,或者需要更多的数据。

③【最大似然(ML)】。最大似然法。使用迭代的方法产生与实际观测的数据最一致的估计,这些估计可能是有偏差的。该方法是接近正态的。ML 法和 REML 法估计在经转换后是不变的。该方法对固定效应作估计时未考虑自由度。

④【约束最大似然法(REML)】。也叫有限最大似然法。对于许多平衡数据(并非对所有平衡数据),该方法比 ANOVA 法估计值要小。因为此方法对固定效应作了调整,计算的标准误可能比 ML 法要小。在估计固定效应时考虑自由度。

(2)【随机效果优先】栏。在系统默认 MINQUE 方法的同时,激活【随机效果优先】栏,也就是说该栏只对 MINQUE 法有效。

①【相等】。指定此项意味着所有随机效应和残差项对观测的影响相等,是 MINQUE 法中系统默认项。

②【零】。指定此项,假设随机效应方差相等且都为零。仅在指定了 MINQUE 法时可以指定此选项。

(3)【平方和】栏。在指定 ANOVA 法的同时,激活该栏。

①【类型 I】。用于分层模型方差成分的迭代,是系统默认的。

②【类型 III】。仅用于 ANOVA 法。

(4)【输出】栏。

① 在指定 ANOVA 法的同时,激活以下两个复选项:

●【平方和】。要求显示平方和。

●【期望均值平方】。要求显示期望均方值。

② 在指定 ML 法或 REML 法时,只激活【输出】栏中的【迭代历史记录】,要求显示迭代过程。

(5)【标准】栏。对判据给定参数。指定 ML 法或 REML 法的同时才能激活该栏。

① 在【收敛性】框中指定收敛标准,下拉列表中以科学计数法列出选择范围“1E-6~1E-10”,表示  $10^{-6} \sim 10^{-10}$ ,共 5 个选项,可以选择其中之一。

② 在【最大迭代】框中可指定最大迭代次数。

#### 4. 【方差成分:保存】对话框

单击【保存】按钮,打开【方差成分:保存】对话框,见图 9-37。该对话框可以将方差成分分析结果作为一个新的数据文件存储到指定的数据文件中,以便进行其他的统计分析时使用。

(1) 指定保存内容。

①【方差成分估计】。保存方差成分估计值。

②【成分共变】。该项只有在选择 ML 法或 REML 法时才被激活。在以下两项中选择其一:

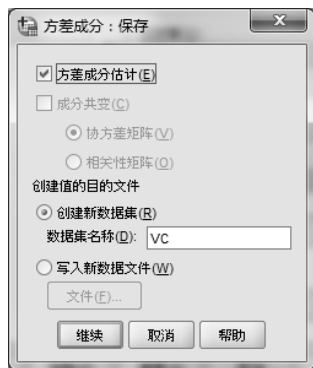


图 9-37 【方差成分:保存】对话框



- 【协方差矩阵】。
- 【相关性矩阵】。

(2) 【创建值的文件】。对所产生的方差成分估计值、和或矩阵指定保存目标。

① 【创建新数据集】。可以保存成一个数据集，但不是外部数据文件，除非事先明确在这样一个 SPSS 期间结束保存成数据文件。数据集只在当前的 SPSS 期间使用，用作其他分析的数据。要在【数据集名称】框中给出数据集的名字。

② 【写入新数据文件】。把分析结果产生的数据写入一个外部数据文件。

选择此项将激活【文件】按钮，单击该按钮，打开【方差成分：保存到文件】对话框，在该对话框中指定保存位置、文件类型和文件名。通常保存类型为“\*.sav”，即 SPSS 数据文件类型。单击【继续】按钮，返回主对话框。

9.6.2 方差成分分析实例

【例 12】 本例使用教育心理学试验中的心理运动测验分数与被试者必须瞄准的目标大小关系的资料，即数据文件 data09-07。

1) 操作步骤

(1) 读取数据文件 data09-07。

(2) 按【分析→一般线性模型→方差分量估计】顺序单击各菜单项，打开【方差成分】主对话框，见图 9-34。

(3) 定义因变量、因素变量和随机因素变量。在【方差成分】主对话框的变量列表中选择得分 score 作为因变量，单击向右箭头，将其送入【因变量】框内。因为被试者瞄准的目标和使用的设备是试验设计者选择的固定条件，所有研究者感兴趣的水平都包括在数据文件中了，属于固定因素。因此，在变量列表内选中 target、device 作为固定因素变量，单击第二个向右箭头按钮，将其送入【固定因子】框中。由于认为亮度是随机地从亮度这个大总体中随机选择的两种亮度，可以认为亮度是随机因素，因此，选择 light 变量作为随机因素，单击第三个向右箭头按钮，将其送入【随机因子】框中。

(4) 单击【模型】按钮，打开【模型选择】对话框，选择【设定】，自定义模型。

① 在【构建项】框下确定【主效应】项，从左边的变量框分别选择 target、device、light 变量，单击向右箭头，送入【模型】框中。这样定义了 3 个主效应。

② 在【构建项】框下选择【交互效应】项，从左边变量框同时选择 target 和 light 变量，单击向右箭头，同时送入【模型】框中，即确定 target\*light 交互项。用同样方法将 device\*light 和 target\*device\*light 送入模型框中。按【继续】按钮，返回【方差成分分析】主对话框。该步骤的目的是对主效应和与随机因素变量有关的交互效应做方差成分估计。

(5) 单击【选项】按钮，打开【选项】对话框。

① 指定方差估计方法。在【模型】栏下，选中【ANOVA】法。

② 在【平方和】栏下选择【类型III】为分解偏差平方和的方法。

③ 在【输出】栏中选择显示【平方和】和【期望均值平方】。

单击【继续】按钮返回主对话框。

2) 运行结果(见表 9-69~表 9-75)

3) 运行结果解释

表 9-69 因素水平情况

		值标签	N
亮度	1	l1	60
	2	l2	60
目标	1	t1	30
	2	t2	30
	3	t3	30
	4	t4	30
设备	1	d1	40
	2	d2	40
	3	d3	40

因变量: score

表 9-69 所示为因素水平情况表，表中列出 3 个因素 target、device、light，每个因素的水平及值标签。表下方注明了因变量是 score。

表 9-70 所示为方差分析结果。表中列出了各主效应和交互效应的平方和分解的结果，即各效应的偏差平方和值、各自的自由度以及均方值。可以看出，此表就是方差分析结果，不管将 light 变量作为固定因素还是随机因素，模型确定时，ANOVA 法的分析结果都是相同的，说明方差成分分析与单因变量多因素分析是兼容的。这里的均方也称作观测均方。

表 9-71 所示为方差成分表，列出了各效应的方差估计值。注解中说明：

表 9-70 方差分析结果

ANOVA			
源	III 型平方和	df	均方
校正的模型	783.467	23	34.064
截距	3162.133	1	3162.133
target	235.200	3	78.400
device	86.467	2	43.233
light	76.800	1	76.800
light * target	93.867	3	31.289
light * device	12.600	2	6.300
light * target * device	278.533	12	23.211
误差	70.400	96	.733
总计	4016.000	120	
校正的总计	853.867	119	

因变量: score

表 9-71 方差成分表

方差估计	
分量	估计
Var(light)	1.040
Var(light * target)	.539
Var(light * device)	-.846 <sup>a</sup>
Var(light * target * device)	4.496
Var(误差)	.733

因变量:score  
方法:ANOVA (III 型平方和)  
a. 对于 ANOVA 和 MINQUE 方法，可能会出现负方差分量估计值。出现负方差分量估计值的可能原因有：(a) 所指定的模型不是正确的模型，或 (b) 方差的真值等于 0。

- ① 因变量为 score。
- ② 分析方法为 ANOVA 法，使用 Type III 方法计算偏差平方和。表中的交互效应项“light\*device”的方差为负值，这是 ANOVA 法可能发生的结果。其原因可能是：

- 指定的模型是错误的。
- 该方差估计的实际值为 0。

亮度与设备的交互效应可以不考虑，因为不但在方差成分表中显示了负值，在方差分析表中的均方值也是最小的，均方值仅为 6.3，见表 9-72。

因此，必须重新设计分析模型，进行第二次分析。只是在前面模型基础上，去掉 light\*device 交互项，结果见表 9-73 和表 9-74。

表 9-72 期望均方系数表

期望均方值						
源	方差分量					
	Var(light)	Var(light * target)	Var(light * device)	Var(light * target * device)	Var(误差)	2 次项
截距	60.000	15.000	20.000	5.000	1.000	截距, target, device target device
target	.000	15.000	.000	5.000	1.000	
device	.000	.000	20.000	5.000	1.000	
light	60.000	15.000	20.000	5.000	1.000	
light * target	.000	15.000	.000	5.000	1.000	
light * device	.000	.000	20.000	5.000	1.000	
light * target * device	.000	.000	.000	5.000	1.000	
误差	.000	.000	.000	.000	1.000	

因变量:score  
期望均方值根据 III 型平方和计算。  
对于每个源，期望均方值等于单元格中系数之和乘以方差分量，再加上与 2 次项单元格中的效应相关的 2 次项。

表 9-73 是第二次方差分析的方差分析表。偏差平方和分解可以与表 9-70 比较，去掉了 light\*device 项，总偏差平方和值不变，该项的偏差平方和包括在“light\*device\*target”中了。

因变量 score 心理测试得分的方差，在交互项中，主要来源于随机变量 light 与固定因素变量 device、target 的三维交互效应。亮度与目标的交互效应不大。可以得出的结论是：亮度对测量得分的影响不能忽视，设备和目标是两个固定因素，是作为试验条件存在的。因此减小方差的方向要从减小它们与亮度的交互效应考虑。

表 9-72 给出了第一次方差分析得出的期望均方与方差成分之间的系数矩阵。方差成分分析的 ANOVA 方法是根据随机效应的期望均方与观测均方相等来估计方差成分的，即根据表 9-72 和表 9-73 得出表 9-74 的结果。

表 9-73 第二次方差分析

ANOVA			
源	III 型平方和	df	均方
校正的模型	783.467	23	34.064
截距	3162.133	1	3162.133
target	235.200	3	78.400
device	86.467	2	43.233
light	76.800	1	76.800
light * target	93.867	3	31.289
light * target * device	291.133	14	20.795
误差	70.400	96	.733
总计	4016.000	120	
校正的总计	853.867	119	

因变量: score

表 9-74 第二次分析的方差成分表

方差估计	
分量	估计
Var(light)	.759
Var(light * target)	.700
Var(light * target * device)	4.012
Var(误差)	.733

因变量:score  
方法:ANOVA (III 型平方和)

本例中，根据表 9-72 可得：  
 $EMS(light*target*device) = 5*Var(light*target*device)+1*Var(Error)$   
 $MS(light*target*device) = 20.795$  (见表 9-73)  
 $Var(Error) = 0.733$  (见表 9-73)

可以解出三阶交互效应的方差成分  $Var(light*target*device)=4.012$ 。

根据表 9-72 可得：  
 $EMS(light*target) = 5* Var(light*target*device)+15*Var (light*target)+ Var(Error)$

可以解出随机效应的二阶交互效应的方差成分  $Var (light*target) = 0.700$ 。

读者可以自己解出随机因素 light 的主效应的方差成分值为 0.759。

下面列出不同分析方法的方差成分分析结果，可以看出，不同方法输出结果在数值上稍有差别，但趋势一致、结论一致。

表 9-74 是方差成分表，可以看出，设计的模型包括除了“light\*device”外所有可能的与亮度变量 light 有关的效应项，其方差估计值的总和为

$$\begin{aligned} Var(light)+Var(light*target)+Var(light*target*device) &= 0.759 + 0.700 + 4.012 = 5.471 \\ \text{light 的各阶效应的总方差估计值/ (light 的各阶效应的总方差估计值+Var(Error))} \\ &= 5.471/(5.471+0.733) = 88.19\%。 \end{aligned}$$

亮度的效应解释了随机效应的 88.19%，误差效应仅解释了随机效应的 11.91%，说明亮度对随机效应的贡献相当多，在该项心理测试试验中是不可忽视的。而在该表中还可以看出，三维交互项所解释的方差是 4.012，在亮度的所有可能的效应中占了 73.3%，因此三维效应又是最值得关注的。

表 9-75 所示是 ML 最大似然法的方差成分分析结果。

与 ANOVA 法一样，均说明方差最大来源于亮度、目标、设备的交互效应。亮度因素是不可忽视的，亮度应该在测试中作为测试条件考虑。

表 9-75 ML 法的方差成分分析结果

方差估计	
分量	估计
Var(light)	.379
Var(light * target)	.264
Var(light * target * device)	2.190
Var(误差)	.733

因变量:score  
方法:极大似然估计

习 题 9

1. 简述方差分析的基本思想。用表达式表示单因素方差分析的偏差平方和分解。
2. 方差分析的假定的前提条件有哪些？
3. 什么是主效应？什么是交互效应？
4. 简述协方差分析的基本思想。
5. 对 4 个服务行业的服务质量进行评价，较高得分表示较高的服务质量。对航空公司、零售业、旅馆业和汽车制造业进行的评定数据见数据文件 data09-12。在显著性水平  $\alpha = 0.05$  下，检验 4 种行业质量等级的总体均值是否差异显著。
6. 数据文件 data09-13 是 474 个银行职工的数据。试分析银行办事员起始工资是否与职工的性别、民族有关。分析时假定银行办事员起始工资总体为正态分布，不考虑其他因素的影响。
7. 数据文件 data09-14 是 15 名手术要求基本相同的患者，随机分为 3 组，分别在手术中使用 3 种麻醉诱导方法 A、B、C，在不同时相(诱导前 T0 和 T1、T2、T3、T4)测量的收缩压数据。试进行方差分析。

# 第10章 相关分析

## 10.1 相关分析的概念与相关分析过程

### 10.1.1 简单相关分析的概念

#### 1. 两个变量间的简单相关分析

相关分析是研究变量间关联密切程度的一种常用统计方法。线性相关分析研究两个变量间线性关系的强弱程度和方向。相关系数是描述线性关系强弱程度和方向的统计量，通常用  $r$  表示。

如果一个变量  $y$  可以确切地用另一个变量  $x$  的线性函数表示，这种关系是确切的，则两个变量间的相关系数是 1 或 -1。

一般情况，两个变量的对应关系不具有唯一性。例如身高与体重的关系，相同身高的人会有不同的体重。研究它们之间线性关系的密切程度使用相关分析。

变量  $y$  随着变量  $x$  的增加而增加或随着变量  $x$  的减少而减少，称为变化方向一致。发育阶段的少年，身高越高，体重相对也就越大。这种相关称为正向相关，其相关系数大于 0。如果变量  $y$  随着变量  $x$  的增加而减少，例如吸烟量和吸烟时间与肺功能的关系，变化方向相反。随着吸烟量增加，肺功能下降；随着吸烟时间加长，肺功能下降。这种相关关系称为负相关，其相关系数小于 0。相关系数  $r$  没有单位，其值在 -1~1 之间。

正态分布变量  $x$  与  $y$  间的线性相关系数采用如下 Pearson 积矩相关公式计算

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

式中， $\bar{x}$ 、 $\bar{y}$  分别是变量  $x$ 、 $y$  的均值； $x_i$ 、 $y_i$  分别是变量  $x$ 、 $y$  的第  $i$  个观测值。

#### 2. 非参相关分析

如果数据分布不满足正态分布的条件，应使用 Spearman 和 Kendall 相关分析方法。

(1) Spearman 相关系数是 Pearson 相关系数的非参形式，是根据数据的秩而不是根据实际值计算的。也就是说，先对原始变量的数据排序，根据各秩使用 Spearman 相关系数公式进行计算。它适合有序数据或不满足正态分布假设的等间隔数据。相关系数值的范围也是在 -1~1 之间，绝对值越大，表明相关性越强。相关系数的符号也表示相关的方向。这两种相关系数的计算必须对变量值排序。变量  $x$ 、 $y$  之间的 Spearman 相关系数计算公式为

$$\theta = \frac{\sum (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum (R_i - \bar{R})^2 \sum (S_i - \bar{S})^2}}$$

式中， $R_i$  是第  $i$  个  $x$  值的秩； $S_i$  是第  $i$  个  $y$  值的秩； $\bar{R}$ 、 $\bar{S}$  分别是  $R_i$  和  $S_i$  的平均值。

(2) Kendall's tau 系数也是一种对两个有序变量或两个秩变量间的关系程度的测度,因此也属于一种非参测度。分析时考虑了结点(秩次相同的)的影响。Kendall's tau 系数的计算公式为

$$\tau = \frac{\sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)}{\sqrt{(T_0 - T_1)(T_0 - T_2)}}$$

式中,  $\text{sgn}(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$ ;  $T_0 = n(n-2)/2$ ;  $T_1 = \sum t_i(t_i-1)/2$ ;  $T_2 = \sum u_i(u_i-1)/2$ ;  $t_i$ (或  $u_i$ )

是  $x$ (或  $y$ ) 的第  $i$  组结点  $x$ (或  $y$ ) 值的数目;  $n$  为观测数。

两个或若干变量之间或两组观测之间的关系,有时也可以用相似性或不相似性来描述。相似性测度用大数值表示很相似,较小的数值表明相似性小。不相似性使用距离或不相似性来描述,大值表示相差甚远,有关内容参见第 10.4 节。

3. 关于相关系数统计意义的检验

由于我们通常是通过抽样方法,利用样本研究总体的特性。由于抽样误差的存在,样本中两个变量间相关系数不为 0,不能说明总体中这两个变量间的相关系数不是 0,因此必须经过检验。检验的零假设是:总体中两个变量间的相关系数为 0。SPSS 的相关分析过程给出了该假设检验的概率, Pearson 和 Spearman 相关系数假设检验  $t$  值计算公式为

$$t = \frac{\sqrt{n-2}r}{\sqrt{1-r^2}}$$

式中,  $r$  是相关系数;  $n$  是样本观测数;  $n-2$  是自由度。当  $t > t_{0.05(n-2)}$  时,  $p < 0.05$ , 拒绝原假设,否则不足以在这个检验中拒绝相关系数为 0 的原假设。相关系数不等于 0 并不意味着相关,只有当它大于某些预期的数目时才能认为相关。

在 SPSS 的相关分析过程的输出中只给出相关系数和假设检验的概率  $p$  值。

10.1.2 相关分析过程

在【分析】菜单中的【相关】命令项有 3 个相关分析功能命令项,见图 10-1。

1. 【双变量】命令项

调用 Correlations 过程和 Nonpar Corr 过程,按指定项显示变量的描述统计量。计算指定的两个变量间的相关系数,可以对应地去选择计算 Pearson 相关系数、Spearman 相关系数和 Kendall's tau-b 相关系数,同时对相关系数进行检验。检验的零假设是:总体中两个变量间的线性相关系数为 0。可以对检验进行单尾或双尾的选择,给出检验相关系数为 0 的概率。

2. 【偏相关】命令项

调用 Partial Corr 过程,计算两个变量间在控制了其他变量的影响下的相关系数。可以选择单尾或双尾显著性检验。检验的零假设是:总体中两个变量间偏相关系数为 0。还可以要求计算其他描述统计量。

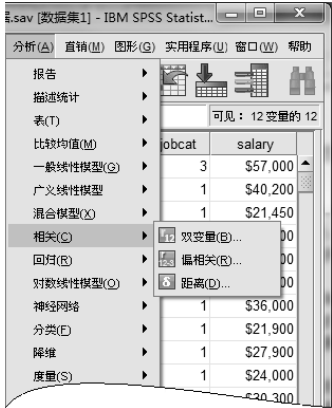


图 10-1 【分析】菜单中的【相关】命令项

3. 【距离】命令项

调用 Proximities 过程，对变量或观测进行相似性或不相似性测度。因此分析的变量可以是连续变量、表示频数分布的变量，某些测度还适用于二值变量。还可以对原始数据和计算出的距离数据进行标准化。

如果为达到预测目的，研究自变量的变动对因变量的影响程度，根据已知自变量的变化来估计因变量的变化情况，必须使用回归分析。

10.2 两个变量间的相关分析

本节介绍两个变量间的相关，包括两个连续变量间的相关和两个等级变量间的秩相关。这两种相关使用同一个命令项 Bivariate 调用。在对话框中，可通过选择不同的分析方法调用不同的分析过程。选择哪一种分析方法要看具体的数据类型。

10.2.1 两个变量间的相关分析过程

在进行相关分析之前，应使用【图形】菜单中的【散点/点状】命令作散点图，进行初步观察，确认两个变量间有相关趋势，再按下列步骤进行相关分析。

1. 选择分析变量

按【分析→相关→双变量】顺序单击菜单项，打开【双变量相关】分析主对话框，见图 10-2。在左边的变量表中选择两个以上变量送入【变量】框中。

2. 相关系数栏中列出相关分析类型

(1) 【Pearson】。皮尔逊相关，系统默认的相关分析方法。只有正态分布的等间隔测度的变量才使用这种相关分析。

(2) 【Kendall 的 tau-b】。肯德尔 $\tau$ -b，调用 Nonpar Corr 非参数相关过程，考虑结点的影响，计算分类变量间的秩相关。

(3) 【Spearman】。斯皮尔曼相关，调用 Nonpar Corr 非参数相关过程计算斯皮尔曼秩相关系数。

如果参与分析的变量是连续变量，选择【Kendall 的 tua-b】或【Spearman】相关，则系统自动对连续变量的值先求秩，再计算其秩分数间的相关系数。

3. 【显著性检验】栏

该栏列出两种显著性检验选项检验针对的零假设是：总体中两个变量不相关。检验结果显示假设检验的概率。

(1) 【双侧检验】。即双尾 T 检验，系统默认的检验方式，当事先不知道相关方向(正相关还是负相关)时选择此项。

(2) 【单侧检验】。单尾 T 检验，如果事先知道相关方向可以选择此项。



图 10-2 【双变量相关】分析主对话框

4. 【标记显著性相关】

选择该项要求在输出结果中，相关系数右上方使用“\*”表示显著性水平为 5%，用“\*\*”表示其显著性水平为 1%。

5. 选择项

在主对话框中单击【选项】按钮，打开如图 10-3 所示对话框。

(1) 【统计量】选项。选择对输出的要求。只有在主对话框中选择了【Pearson】相关分析方法才可以选择这两个选项。

- ① 【均值和标准差】。输出两个描述统计量。
- ② 【叉积偏差和协方差】。输出叉积离差矩阵和协方差矩阵。

(2) 【缺失值】栏。在该栏中选择缺失值处理方法。

① 【按对排除个案】。仅剔除正在参与计算的两个变量值都是缺失值的观测。这样，有可能在计算出的相关系数矩阵中，相关系数是根据不同数量的观测计算出来的。选择此项，可以最大限度地使用取得的观测数据。

② 【按列表排除个案】。剔除在主对话框【变量】栏中列出的变量带有缺失值的所有观测。输出的相关矩阵中，每个相关系数都是依据相同数量的观测计算出来的。



图 10-3 【双变量相关性: 选项】对话框

10.2.2 两个变量间的相关分析实例

【例 1】 使用默认选项进行简单相关分析的例题。

(1) 使用数据文件 data10-01，以 1962—1988 年安徽省“国民收入”与“城乡居民储蓄存款余额”两个变量间的线性相关分析为例，说明使用系统默认值进行连续变量相关分析的方法。数据来源于中国现场统计研究会主办的《数理统计与管理》1990 年第 5 期。变量包括：income(国民收入(亿元))、deposit(城乡居民储蓄存款余额)、number(序号)、year(年份)。

(2) 操作说明：读取数据文件 data10-01，在源变量栏中选择分析变量 deposit(城乡居民储蓄存款余额)和 income(国民收入)，单击向右箭头按钮，将选择的变量移至【变量】框中。其余使用系统默认选项。单击【确定】按钮提交系统执行。输出结果见表 10-1。

表 10-1 安徽省国民收入与城乡居民存款储蓄余额的相关分析

相关性		国民收入(亿元)	城乡居民储蓄存款余额
国民收入(亿元)	Pearson 相关性	1	.976**
	显著性(双侧)		.000
	N	27	27
城乡居民储蓄存款余额	Pearson 相关性	.976**	1
	显著性(双侧)	.000	
	N	27	27

\*\* . 在 .01 水平(双侧)上显著相关。

表 10-1 所示是安徽省国民收入变量 income 和城乡居民存款余额变量 deposit 之间的相关系数矩阵，在变量行与变量列的交叉处纵向显示了 3 个数值。

第一行中的数值是行变量与列变量的相关系数矩阵。行、列变量相同，其相关系数为 1。变量国民收入与城乡居民储蓄存款余额之间的相关系数为 0.976。

第二行中的数值是使相关系数为 0 的假设检验成立的概率，结果均小于 0.001。



第三行中的数值是参与该相关系数计算的观测数目，均为 27。

注释行说明标有 “\*\*” 的相关系数的显著性概率水平为 0.01。由于  $r = 0.976$ ，因此，国民收入与存款余额之间是高度相关的。

【例 2】 生成矩形相关矩阵的简单相关例题。

(1) 使用数据文件 data10-02，为一组银行雇员数据。分析的目的是要观察 salary(当前工资)与 salbegin(起始工资)、雇员本人各方面条件的关系。变量有：salary(当前工资)、salbegin(起始工资)、age(年龄)、jobtime(以月为本单位的工作时间)、prevexp(以月为本单位的以前工作经历)。

(2) 操作步骤如下：

- ① 读取数据文件 data10-02，按【分析→相关→双变量】顺序单击菜单项，打开相应的对话框。
- ② 在源变量框中选择 jobtime、prevexp、age、salary、salbegin 送入【变量】框作为分析变量。
- ③ 主对话框中的选项。
  - 分析方法选择【Person】相关。
  - 【显著性检验】栏中选择【双侧检验】。
  - 选中【标记显著性相关】复选项。
- ④ 在【双变量：选项】对话框中指定。
  - 【统计量】栏中选择【均值和标准差】。
  - 【缺失值】栏中选择【按对排除个案】，按对剔除带有缺失值的观测。

单击【继续】按钮返回主对话框，单击【确定】按钮提交运行。

⑤ 运行程序语句如下：

```
CORRELATIONS                                ①
/VARIABLES= salary salbegin jobtime prevexp age  ②
/PRINT=TWOTAIL NOSIG                          ③
/STATISTICS DESCRIPTIVES                       ④
/MISSING=PAIRWISE .                             ⑤
```

得出的描述统计量见表 10-2。相关矩阵在表 10-3(a)中生成程序后进行修改，由于只需要变量 salary 与其他各变量的相关性，因此在第②语句 salary 与其他变量之间增加 “with”，以便使结果更加清晰，即

```
/VARIABLES= salarywith salbegin jobtime
age prevexp.
```

- (3) 程序运行结果见表 10-2 和表 10-3(b)。
- (4) 结果分析。

表 10-2 所示是相关分析变量的描述统计量。可以看出，当前工资的平均值比起始工资要高，而且当前工资标准差比起始工资标准差大了，说明当前工资差别大了。

表 10-3(a)是根据对话框指定的选项运行的结果。显然表格表达得过于烦琐，没有仅针对题目要求显示当前工资与其他变量的相关。

表 10-3(b)是经过修改的程序运行的结果，简单明了。

在行变量与列变量的交叉点单元格上，第一行为 Pearson 相关性，即皮尔逊相关系数；第二行表中 Sig. (2-tailed) 是对于相关系数为 0 的假设的双尾 T 检验结果，它是出现目前统计量值及其更加极端值的概率  $p$ ；第三行 N 为参与相关系数计算的有效观测数。

表 10-2 分析变量的描述统计量

描述性统计量			
	均值	标准差	N
受雇月数	81.11	10.061	474
过去经验(月)	95.86	104.586	474
年龄	47.14	11.775	473
当前工资	\$34,419.57	\$17,075.661	474
起始工资	\$17,016.09	\$7,870.638	474

表 10-3(a) 相关矩阵表

		相关性				
		受雇月数	过去经验(月)	年龄	当前工资	起始工资
受雇月数	Pearson 相关性	1	.003	.054	.084	-.020
	显著性 (双侧)		.948	.244	.067	.668
	N	474	474	473	474	474
过去经验(月)	Pearson 相关性	.003	1	.802**	-.097*	.045
	显著性 (双侧)	.948		.000	.034	.327
	N	474	474	473	474	474
年龄	Pearson 相关性	.054	.802**	1	-.144**	-.010
	显著性 (双侧)	.244	.000		.002	.833
	N	473	473	473	473	473
当前工资	Pearson 相关性	.084	-.097*	-.144**	1	.880**
	显著性 (双侧)	.067	.034	.002		.000
	N	474	474	473	474	474
起始工资	Pearson 相关性	-.020	.045	-.010	.880**	1
	显著性 (双侧)	.668	.327	.833	.000	
	N	474	474	473	474	474

\*\* 在 .01 水平 (双侧) 上显著相关。

\* 在 0.05 水平 (双侧) 上显著相关。

表 10-3(b) 简单明了的相关矩阵

		相关性			
		起始工资	年龄	受雇月数	过去经验(月)
当前工资	Pearson 相关性	.880**	-.144**	.084	-.097*
	显著性 (双侧)	.000	.002	.067	.034
	N	474	473	474	474

\*\* 在 .01 水平 (双侧) 上显著相关。

\* 在 0.05 水平 (双侧) 上显著相关。

很明显,“当前工资”与“起始工资”相关系数最大,为 0.88,相关系数为 0 的概率小于 0.001,因此,“当前工资”与“起始工资”之间有高度正相关。虽然“年龄”、“过去经验”及“受雇月数”与“当前工资”在有关相关系数为零的检验中,检验的概率值也较小。但其相关系数值均较小,故不能认为它们之间存在线性相关。

【例 3】 秩相关实例。

- (1) 使用数据文件 data10-02。说明: 以上对雇员的工资的分析并不严格,因为虽然参与分析的变量均为尺度(连续)变量,但没有作各变量是否符合正态分布的检验。下面使用秩相关分析方法分析各雇员的 salary(当前工资)与 salbegin(起始工资)、educ(受教育程度)、过去经验(prevexp)、受雇月数间的关系。educ 数值数小于 24(系统参数定义的),因此属于有序分类变量。
- (2) 重新启动双变量相关分析。移入【变量】框中的变量有 salbegin、salary、educ、prevexp、jobtime。分析方法选择【Kandall 的 tau-b】,即秩相关;选择【双侧检验】;选中【标记显著性相关】复选项。【缺失值】处理方法选择【按对排除个案】。
- (3) 运行程序语句。在主对话框中单击【粘贴】按钮,在语句窗口中生成如下程序:

```
NONPAR CORR /VARIABLES= salary salbegin educ prevexp jobtime
/PRINT=KENDALL TWOTAIL NOSIG
/MISSING=PAIRWISE .
```

生成程序后对语句进行修改,修改 VARIABLES 语句为如下形式:

```
/VARIABLES= salary with salbegin educ prevexp jobtime
```

(4) 输出结果见表 10-4。

表 10-4 非参相关矩阵

相关系数			salbegin 起始 工资	educ 受教育程 度(年)	prevexp 过去 经验(月)	jobtime 受雇 月数
Kendall 的 tau_b	salary 当前工资	相关系数	.656**	.554**	-.013	.071
		Sig. (双侧)	.000	.000	.677	.022
		N	474	474	474	474

\*\* 在置信度 (双侧) 为 0.01 时, 相关性是显著的。

\* 在置信度 (双侧) 为 0.05 时, 相关性是显著的。

(5) 输出结果分析。表 10-4 中, “当前工资”与“起始工资”秩相关系数值较大, 为 0.656, 与“受教育程度”的秩相关系数为 0.554, 相关系数为零的概率几乎为 0, 说明“当前工资”与“起始工资”存在中度的秩相关。而“当前工资”与“受雇月数”及“过去经验”的相关系数很小, 因而不能说明它们之间存在秩相关。

读者可以根据自己的经验分析这样的工资结构是否合理, 是否有利于调动职工的积极性, 从而有利于企业的发展。

【例 4】两个等级变量间的秩相关实例。

数据文件 data10-03 为某次全国武术比赛女子前 10 名运动员长拳和长兵器两项得分的数据, 要求分析这两项得分是否在线性关系。变量 score1、score2 分别为长拳和长兵器两项得分, 变量 ranking 为名次。

(1) 读取数据文件 data10-03。

(2) 按【分析→相关→双变量】顺序单击菜单项, 打开【双变量相关】分析主对话框。移入【变量】框中的变量为 score1、score2; 分析方法选择【Kendall 的 tau-b】、【Spearman】; 显著性检验类型选择【单侧检验】; 选择【标记显著性相关】复选项。

(3) 在【双变量相关性: 选项】对话框中, 选择【按对排除个案】, 即成对剔除带有缺失值的观测。改变 Varibles 子命令, 在两个变量之间加“with”。

(4) 运行的程序如下:

```
NONPAR CORR /VARIABLES=score1 with score2
               /PRINT=BOTH ONETAILED NOSIG /MISSING=PAIRWISE .
```

(5) 输出结果见表 10-5。从表中可以看到, Kendall 的 tau-b 相关系数为 0.543, 单尾检验的概率为 0.027, 小于 0.05。Spearman 相关系数是 0.610, 单侧检验的概率为  $p = 0.030$ , 小于 0.05, 由于样本量较小, 故两项得分间线性相关程度不高。

对于非等间隔测度的连续变量, 因为分布不明, 可以使用等级相关分析, 如上一个例 3 也可以使用 Pearson 相关分析; 对于完全等间隔的离散变量, 则必须使用等级相关分析相关性。

表 10-5 Kendall's tau-b 与 Spearman 相关系数

相关系数			长拳得分
Kendall 的 tau_b	长兵器得分	相关系数	.543
		Sig. (单侧)	.027
		N	10
Spearman 的 rho	长兵器得分	相关系数	.610
		Sig. (单侧)	.030
		N	10

\* 在置信度 (单侧) 为 0.05 时, 相关性是显著的。

## 10.3 偏相关分析

### 10.3.1 偏相关分析的概念

#### 1. 偏相关分析

简单相关分析计算两个变量间的相关系数，分析两个变量间线性关系的程度和方向。往往因为第三个变量的作用，使相关系数不能真正反映两个变量间的线性程度，如身高、体重与肺活量之间的关系。如果使用 Pearson 相关分析计算其相关系数，可以得出肺活量与身高和体重均存在较强的线性关系。但实际上，如果对体重相同的人，是否身高值越大，肺活量越大呢？结论是否定的。因为身高与体重有着线性关系，体重与肺活量存在线性关系，从而得出身高与肺活量之间存在较强的线性关系的结论是错误的。偏相关分析的任务就是在研究两个变量之间的线性相关关系时控制可能对其产生影响的变量。分析身高与肺活量之间的相关性，就要控制体重在相关分析中的影响。实际生活中有许多这样的关系，例如，可以控制年龄和工作经验两个变量的影响，估计工资收入与受教育程度之间的相关关系；可以在控制销售能力与各种其他经济指标的情况下，研究销售量与广告费用之间的关系等。

#### 2. 偏相关系数的计算

控制了变量  $z$ ，变量  $x$ 、 $y$  之间的偏相关和控制了两个变量  $z_1$ 、 $z_2$ ，变量  $x$ 、 $y$  之间的偏相关系数计算公式如下：

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$
$$r_{xy,z_1z_2} = \frac{r_{xy,z_1} - r_{xz_2,z_1}r_{yz_2,z_1}}{\sqrt{(1-r_{xz_2,z_1}^2)(1-r_{yz_2,z_1}^2)}}$$

第一个公式中的  $r_{xy,z}$  是控制了  $z$  的条件下， $x$ 、 $y$  之间的偏相关系数。 $r_{xy}$  是变量  $x$ 、 $y$  间的简单相关系数或称零阶相关系数； $r_{xz}$ 、 $r_{yz}$  分别是变量  $x$ 、 $z$  间的和变量  $y$ 、 $z$  间的简单相关系数，依此类推。

#### 3. 偏相关系数的检验

在利用样本研究总体的特性时，由于抽样误差的存在，样本中控制了其他变量的影响，两个变量间偏相关系数不为 0，不能说明总体中这两个变量间的偏相关系数不是 0，因此必须进行检验。检验的零假设：总体中两个变量间的偏相关系数为 0。使用 T 检验方法，公式如下：

$$t = \frac{\sqrt{n-k-2}r}{\sqrt{1-r^2}}$$

这是对 Pearson 偏相关系数假设检验的  $t$  统计量的计算公式。式中， $r$  是相应的偏相关系数； $n$  是观测数； $k$  是控制变量的数目； $n-k-2$  是自由度。当  $t > t_{0.05(n-k-2)}$  时， $p < 0.05$ ，拒绝原假设，否则不足以在这个检验中拒绝变量间偏相关系数为 0 的零假设。

在 SPSS 的偏相关分析过程的输出中只给出偏相关系数和假设检验的概率  $p$  值。

10.3.2 偏相关分析过程

1. 选择分析变量

(1) 按【分析→相关→偏相关】顺序单击菜单项，打开如图 10-6 所示的【偏相关】分析主对话框。

(2) 从左边的变量表中选择分析变量送入【变量】框中；选择控制变量送入【控制】框中。

2. 在【显著性检验】栏

选择假设检验类型，有两个选项：

(1) 【双侧检验】。用于有正、负相关两种可能的情况，是系统默认方式。

(2) 【单侧检验】。用于只可能是正向或只可能是负向相关的情况。

3. 是否显示实际的显著性水平

选择【显示实际显著性水平】复选项，在显示相关系数的同时，显示实际的显著性概率；不选择此项，其显著性概率使用星号“\*”代替，表示其显著性概率在 5%~1%之间，“\*\*”表示其显著性概率小于或等于 1%。

4. 【偏相关性：选项】对话框中的选项

在主对话框中单击【选项】按钮，打开如图 10-7 所示的对话框。



图 10-6 【偏相关】分析主对话框



图 10-7 【偏相关性：选项】对话框

(1) 【统计量】栏

- 【均值和标准差】。要求计算并显示各分析变量的均值和标准差。
- 【零阶相关系数】。要求显示零阶相关矩阵，即 Pearson 相关矩阵。

(2) 【缺失值】栏

- 【按列表排除个案】。剔除所有带有缺失值的观测。系统默认为此项。
- 【按对排除个案】。成对剔除带有缺失值的观测。

选择完成后，单击【继续】按钮返回主对话框，单击【确定】按钮提交系统执行。

10.3.3 偏相关分析实例

【例 5】 使用四川绵阳地区 3 年生中山柏的数据，分析月生长期与月平均气温、月降雨量、月平均日照时数、月平均湿度这 4 个气候因素哪个因素有关。数据来源于袁佳祖编著《灰色系统理论》，数据文件为 data10-04。

这 4 个气候因素彼此均有影响,分析时应对生长量与 4 个气候因素分别求偏相关,在求生长量与 1 个气候因素的相关时控制其他因素的影响,然后比较相关系数,按 4 个气候因素对中山柏生长量影响的大小排序。

(1) 定义变量: month(月份)、hgrow(生长量, cm)、temp(月平均气温, °C)、rain(月降雨量, mm)、hsun(月平均日照时数)、humi(月平均湿度)。输入数据和求简单相关系数的操作略去。

(2) 按【分析→相关→偏相关】顺序单击菜单项,打开【偏相关】分析主对话框。

(3) 指定分析变量和控制变量。

为操作简便,首先确定第一次分析的变量和控制变量。第一次分析变量是生长量(hgrow)与月平均日照时数(hsun),控制变量是月平均湿度(humi)、降雨量(rain)、月平均气温(temp)个变量。

(4) 指定选项。

① 主对话框中的选项使用系统默认值,即选择【双侧检验】、【显示实际显著性水平】。

② 在【偏相关性:选项】对话框中不选择任何选项。假定已经对各变量进行过探索分析,不要各变量的描述统计量。为对比,单写一段计算 Pearson 相关的程序,以简化相关矩阵(见下第一段程序),对缺失值的处理使用系统默认的方法。

(5) 在主对话框中,单击【粘贴】按钮。在语句窗口中生成第一次分析的程序:

PARTIAL CORR	①
/VARIABLES= hgrow hsun BY humi rain temp	②
/SIGNIFICANCE=TWOTAIL	③
/MISSING=LISTWISE	④

程序解释:

① PARTIAL CORR 语句调用偏相关分析过程。

② VARIABLES 子命令定义分析变量与控制变量。BY 前面的 hgrow hsun 为要求相关系数的分析变量,BY 后面的是 humi、rain、temp 这 3 个控制变量。

③ SIGNIFICANCE 子命令要求进行双尾显著性检验。

④ MISSING 子命令要求剔除所有带有缺失值的观测。

此程序即下面的第二段程序(以一个英文句号作为一段程序的结束标志)。

(6) 复制与修改。

① 在【语法】对话框中,选择第一次偏相关分析程序,复制并粘贴 3 次。

② 修改各复制的程序中的 VARIABLES 子命令,改变分析变量和控制变量,形成下面的第三、四、五段程序(第一段程序是求生长量变量与其他气候变量的 Pearson 相关系数,作为对比用):

CORRELATIONS	
/VARIABLES=hgrow with hsun humi rain temp	
/PRINT=TWOTAIL NOSIG	第一段
/MISSING=PAIRWISE .	
PARTIAL CORR	
/VARIABLES= hgrow hsun BY humi rain temp	第二段
/SIGNIFICANCE=TWOTAIL	
/MISSING=LISTWISE .	
PARTIAL CORR	
/VARIABLES= hgrow humi BY hsun rain temp	第三段

```

/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
PARTIAL CORR
/VARIABLES= hgrow rain BY hsun humi temp
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
PARTIAL CORR
/ hgrow temp BY hsun humi rain
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
```

(7) 执行以上各段程序。在输出窗口中显示的部分结果见表 10-6、表 10-7。

表 10-6 生长量与各变量间 Pearson 相关分析结果

		相关性			
		temp 月平均 气温(c)	rain 月降雨量 (mm)	hsun 月平均 日照时数	humi 月平均 湿度
hgrow 生长量(cm)	Pearson 相关性	.983**	.709**	.704*	.374
	显著性 (双侧)	.000	.010	.011	.232
	N	12	12	12	12

\*\* 在 .01 水平 (双侧) 上显著相关。  
\* 在 .05 水平 (双侧) 上显著相关。

(8) 分析结果解释与结论。

从表 10-6 的零阶相关矩阵可以看出,“生长量”与“月平均湿度”的相关系数最小,显著性检验结果是不相关的概率为 0.232。结论是生长量除与月平均湿度无关外,与其他几个气候因素均有明显的线性关系。

由于各气候因素的相互影响,例如月平均日照时数与月平均气温高度相关。生长量与各变量间的相关系数并未反映出各变量间的真实情况,因此应该看偏相关的结果。

根据生长量与各气候因素单独的偏相关分析结果,偏相关系数见表 10-7。

表 10-7 偏相关分析结果

相关性				
控制变量			hgrow 生长量 (cm)	hsun 月平均 日照时数
temp 月平均气温(c) & rain 月降雨量(mm) & humi 月 平均湿度	hgrow 生长量(cm)	相关性	1.000	.632
		显著性 (双侧)	.	.068
		df	0	7
	hsun 月平均日照时数	相关性	.632	1.000
		显著性 (双侧)	.068	.
		df	7	0

相关性				
控制变量			hgrow 生长量 (cm)	temp 月平均 气温(c)
rain 月降雨量(mm) & humi 月平均湿度 & hsun 月平均日照时数	hgrow 生长量(cm)	相关性	1.000	.977
		显著性 (双侧)	.	.000
		df	0	7
	temp 月平均气温(c)	相关性	.977	1.000
		显著性 (双侧)	.000	.
		df	7	0

(续表)

相关性			hgrow 生长量 (cm)	rain 月降雨量 (mm)
控制变量				
temp 月平均气温(c) & humi 月平均湿度 & hsun 月平均日照时数	hgrow 生长量(cm)	相关性	1.000	-.491
		显著性 (双侧)	.	.180
		df	0	7
	rain 月降雨量(mm)	相关性	-.491	1.000
		显著性 (双侧)	.180	.
		df	7	0

相关性			hgrow 生长量 (cm)	humi 月平均 湿度
控制变量				
temp 月平均气温(c) & rain 月降雨量(mm) & hsun 月 平均日照时数	hgrow 生长量(cm)	相关性	1.000	.731
		显著性 (双侧)	.	.025
		df	0	7
	humi 月平均湿度	相关性	.731	1.000
		显著性 (双侧)	.025	.
		df	7	0

根据表 10-7 可总结出表 10-8。根据表 10-8 可以得出结论：中山柏“生长量”与“月均气温”的关系最密切，相关系数为 0.977，不相关的概率  $p < 0.001$ ；其次是“月均湿度”，相关系数为 0.731，假设检验的概率为 0.025；与“月均日照时数”的相关系数为 0.632，不相关的概率为 0.068；与“月均降雨量”的相关系数是负值，但无统计意义，降雨量过大，会影响其生长。可以看出，偏相关分析结果与简单相关分析结果会有很大区别。

表 10-8 中山柏生长量与四个气候因素的偏相关综合结果

	月均气温	月均湿度	月均日照时数	月降雨量
生长量	.977	.731	.632	-.491
自由度	( 7)	( 7)	( 7)	( 7)
不相关概率 p	0.000	0.025	0.068	0.180

10.4 距离分析

10.4.1 距离分析的概念

1. 关于距离分析

距离分析是对观测之间或变量之间相似或不相似程度的一种测度，是计算一对变量之间或一对观测之间的广义距离。这些相似性或距离测度可以用于其他分析过程，如因子分析、聚类分析或多维定标分析等，有助于分析复杂的数据集。例如，是否可以根据一些特性，如发动机的大小、MPG (每加仑汽油所能行驶的距离) 和马力来测度两种汽车的相似性？通过计算汽车间的相似性，可以对这些汽车获得一些认识，哪些汽车彼此类似，哪些汽车彼此不同。更正规的分析，可以考虑对相似性使用分层聚类或多元定标分析去探测深层结构。

2. 有关的统计量

(1) 不相似性测度

① 对等间隔数据的不相似性(距离)测度可以使用的统计量有：Euclidean 距离(欧几里得



(欧氏)距离)、平方 Euclidean 距离(欧氏距离平方)、Chebychev 距离(切比雪夫距离)、块、Minkowski 距离(明可斯基距离)或设定距离(即自定义统计量)。

② 对计数数据,使用卡方统计量度量或 phi 平方统计量(斐方  $\Phi^2$ )。

③ 对二分类(二值)变量数据,使用 Euclidean 距离(欧氏距离)、平方 Euclidean 距离(欧氏距离平方)、尺寸差分、模式差别、方差、形状或 Lance 和 Williams(兰斯和威廉斯距离)统计量。这些统计量的计算方法可参见第 13.3.3 节关于聚类方法选项(Method)的相关内容。

## (2) 相似性测度

① 等间隔数据使用统计量皮尔逊相关或余弦。

② 测度二元数据相似性使用的统计量有 20 余种,算法参见附录 A。

SPSS 中的距离分析属于专业统计分析过程(Professional Statistics Options),是可选件。如果没有安装,则在菜单中不会有调用该过程的菜单项。

距离分析分为观测之间距离的分析和变量之间距离的分析两类。

## 10.4.2 距离分析过程

SPSS 的距离分析过程提供相似性和不相似性两种分析方法。

### 1. 命令调用

按【分析→相关→距离】顺序单击菜单项,打开【距离】分析主对话框,见图 10-8。其中,【计算距离】栏中,系统默认【个案间】选项;【度量标准】栏中,系统默认【不相似性】方法,即观测间的不相似性测度。在【度量】按钮旁边显示的【Euclidean 距离】表明使用欧几里德(欧氏)距离测度观测间的不相似性。

### 2. 指定分析变量和标识变量

对于观测间的距离分析至少指定一个分析变量和一个标识变量。在源变量栏中选择分析变量,将其移至【变量】框中,选择一个标识变量,将其移至下面一个【标注个案】栏中,见图 10-8。

### 3. 主对话框中的选项

(1) 【计算距离】栏。

① 【个案间】。计算每对观测间的距离。

② 【变量间】。计算每对变量间的距离。

(2) 在【度量标准】栏中选择测度距离的类型与方法。

① 【不相似性】。计算不相似性矩阵,此为系统默认的类型。系统默认使用欧氏距离测度其不相似性。

② 【相似性】。计算相似性矩阵。系统默认使用 Pearson 相关进行相似性测度。

在【度量标准】栏中选择了一种测度类型后,系统默认的计算方法显示在【度量】按钮右侧。可以单击【度量】按钮打开相应的对话框,进一步选择计算方法或统计量。返回主对话框后,被选中的计算方法显示在【度量】按钮旁边。

单击【度量】按钮,打开如图 10-9 所示对话框,指定不相似性测度的计算方法选项。

### 4. 不相似性测度的选项

(1) 【度量标准】栏。选择一种测度需要首先选择数据类型,然后在选中的数据类型组的

下拉菜单中选择与数据类型一致的可用的测度方法，对话框见图 10-9。数据类型及其可以使用的测度如下：

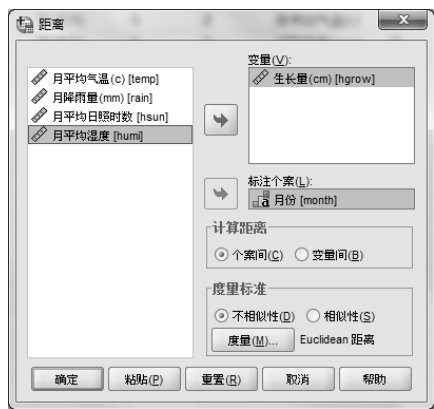


图 10-8 【距离】分析主对话框



图 10-9 【距离：非相似性度量】对话框

①【区间】。等间隔变量(即指连续变量)选项，各选项的详细说明见第 11.3.2 节中的有关内容。

②【计数】。计数变量选项。选择该项后，可以在展开的下拉列表中选择非相似性测度。各选项的详细说明见第 14.3.2 节中的有关内容。

③【二分类】。二值变量(表示某种特性有、无的变量)选项。选择该项后可激活其下的其他选项。在【存在】框中输入表明特性存在的变量值，在【不存在】框中输入表明不存在某特性的变量值。系统默认的变量值是用 1 表明特性存在，用 0 表明特性不存在。对于二值变量的各选项的详细说明参见附录 A 中的有关内容。

(2)【转换值】栏。该栏允许在进行近似计算之前对观测或变量进行标准化，但对二元变量不能进行标准化。

①【标准化】框中可选择标准化的方法，各选项的详细说明见附录。

② 以上除了选项【无】以外，选择其他任意一种标准化的方法，均应同时指定标准化对象，共有两个选项：

- 【按照变量】。即对变量进行标准化。
- 【按照个案】。即对观测进行标准化。

(3)【转换度量】栏。该组选项选择在距离测度计算完成后，对距离测度的结果进行转换的方法。共有 3 种方法，可以同时选择。

①【绝对值】。即对距离取绝对值。当符号表明的是相关的方向且仅对相关的数值感兴趣时使用这种转换。

②【更改符号】。把相似性测度值转换成不相似性测度值或相反。使用这种转换，通过加负号颠倒距离测度的顺序。

③【重新标度到 0-1 全距】。即先减去最小值，然后除以范围(最大值减最小值)使距离标准化。对已经按有意义的方法标准化的测度，一般不再使用此方法进行转换。

5. 相似性测度的选项

在主对话框的【度量标准】栏选择【相似性】选项，单击【度量】按钮，打开【距离：相似性度量】对话框，见图 10-10。

(1) 【度量标准】栏选择相似性测度方法。有关相似性测度方法的选项与非相似性测度一样，在选择具体的测度方法之前必须首先选择变量类型，然后，在选中的实际类型组的下拉菜单中选择与实际类型一致的可用的测度。进行相似性测度的数据类型只有两种：等间隔变量和二元变量。与这两个类型相应的可以选择的测度方法如下：

①【区间】。是等间隔变量选项，各选项的详细说明见附录 A。

②【二分类】。对二元数据的相似性测度。SPSS 为每对项目构造一个 2×2 的列联表。可用的测度落入以下 4 类中：匹配系数、条件概率、可预测性测度和其他测度。可以从下拉列表中选择一种测度。在进行测度方法选择之前应指定表明某特点存在和不存在的变量值。系统默认：特点存在，其值为 1；特点不存在，其值为 0。

读者可以指定其他整数表明特性的出现与不出现，SPSS 将忽略其他值。各选项的详细说明见附录 A。

- (2) 【转换值】栏。转换数值，同相似性测度的对应选项。
- (3) 【转换度量】栏。转换测度，同相似性测度的对应选项。



图 10-10 【距离：相似性度量】对话框

10.4.3 距离分析实例

【例 6】 观测间的相似性分析例题。

仍使用数据文件 data10-04，四川绵阳地区中山柏生长的数据。分析不同月份间生长量之间的距离，以便分析各月份生长量间的相似或不相似性。读入数据文件，按下述步骤操作。

(1) 按【分析→相关→距离】顺序单击菜单项，打开【距离】分析主对话框，见图 10-8 所示。

(2) 指定分析变量和标识变量。选择月生长量 hgrow 作为分析变量，将其移至【变量】框中。选择月份 month 作为标识变量，将其移至下面一个【标注个案】框中。其他使用默认值。

(3) 单击【确定】按钮，提交运行。输出结果见表 10-9 和表 10-10。

表 10-9 观测统计处理简明表  
案例处理摘要

案例					
有效		缺失		合计	
N	百分比	N	百分比	N	百分比
12	100.0%	0	0.0%	12	100.0%

输出结果见表 10-9 和表 10-10。

表 10-9 所示是对观测有效值和缺失值进行的统计。

表 10-10 以矩阵形式给出了两两观测间变量 hgrow 的欧氏距离，即每两个月份间的中山柏生长量间的差值，这是不相似矩阵，行列之间数值越大的不相似性越强。显然，1 月与 8 月生长量最不相似，其欧氏距离值为 19.290；1 月、2 月生长量不相似性最小，其欧氏距离值为 0.490；12 月、2 月、3 月的生长量不相似性和 4 月、10 月的生长量不相似性仅次于 1 月、2 月，其欧氏距离值为 0.5。

在进行观测间不相似性分析时，可以指定若干个分析变量，即根据指定变量组分析观测间的不相似性，标识变量只能指定一个。

【例 7】 变量间的不相似性例题。

对于连续变量间的相似性计算，往往使用 Pearson 相关，这与两个变量间的简单相关分析没有区别。本例仍使用数据文件 data010-04，比较相似性与不相似性的结果。

表 10-10 观测间的欧氏距离

	近似矩阵											
	Euclidean 距离											
	1: 1	2: 2	3: 3	4: 4	5: 5	6: 6	7: 7	8: 8	9: 9	10:10	11:11	12:12
1: 1	.000	.490	1.490	10.790	12.990	16.290	17.990	19.290	14.790	10.290	7.990	.990
2: 2	.490	.000	1.000	10.300	12.500	15.800	17.500	18.800	14.300	9.800	7.500	.500
3: 3	1.490	1.000	.000	9.300	11.500	14.800	16.500	17.800	13.300	8.800	6.500	.500
4: 4	10.790	10.300	9.300	.000	2.200	5.500	7.200	8.500	4.000	.500	2.800	9.800
5: 5	12.990	12.500	11.500	2.200	.000	3.300	5.000	6.300	1.800	2.700	5.000	12.000
6: 6	16.290	15.800	14.800	5.500	3.300	.000	1.700	3.000	1.500	6.000	8.300	15.300
7: 7	17.990	17.500	16.500	7.200	5.000	1.700	.000	1.300	3.200	7.700	10.000	17.000
8: 8	19.290	18.800	17.800	8.500	6.300	3.000	1.300	.000	4.500	9.000	11.300	18.300
9: 9	14.790	14.300	13.300	4.000	1.800	1.500	3.200	4.500	.000	4.500	6.800	13.800
10:10	10.290	9.800	8.800	.500	2.700	6.000	7.700	9.000	4.500	.000	2.300	9.300
11:11	7.990	7.500	6.500	2.800	5.000	8.300	10.000	11.300	6.800	2.300	.000	7.000
12:12	.990	.500	.500	9.800	12.000	15.300	17.000	18.300	13.800	9.300	7.000	.000

这是一个不相似性矩阵

- (1) 按【分析→相关→距离】顺序单击菜单项，打开【距离】分析主对话框。
- (2) 指定分析变量：月平均的气温 temp、降雨量 rain、日照时间 hsun、湿度 humi。选择它们并将其移至【变量】框中。
- (3) 在【计算距离】栏中选择【变量间】，在【度量标准】栏中选择【不相似性】，要求进行变量间的不相似性分析。
- (4) 单击【度量】按钮。

① 在【距离：非相似性度量】对话框中选择【区间】项，因为所选择的变量均为等间隔测度的变量。在下拉列表中选择【Euclidean 距离】，因为只有相似性分析的菜单中才有相关分析的选项，而不相似性分析不计算相关矩阵，只计算欧氏距离或其他距离。

② 因为所选择的分析变量测度的单位不同，因此要对变量进行标准化。在下拉列表中选择【Z得分】，在【转换值】栏中选择【按照变量】项，对变量进行均值为 0，标准差为 1 的标准化。
- (5) 单击【继续】按钮返回主对话框，单击【确定】按钮，提交执行。运行结果见表 10-11。

表 10-11 变量间的不相似性测度标准化后的欧氏距离

	近似矩阵			
	Euclidean 距离			
	月平均气温(c)	月降雨量(mm)	月平均日照时数	月平均湿度
月平均气温(c)	.000	2.505	2.609	3.947
月降雨量(mm)	2.505	.000	2.561	3.680
月平均日照时数	2.609	2.561	.000	4.808
月平均湿度	3.947	3.680	4.808	.000

这是一个不相似性矩阵

如果在主对话框的【度量标准】栏中改选【相似性】选项，则在【距离：非相似性度量】对话框中，在【度量标准】栏中选择【区间】，【度量】选项选择【Pearson 相关性】选项，进行相似性测度分析。运行结果见表 10-12。

比较两种分析结果，可以看出结果是一致的。

表 10-11 所示不相似性分析的结果，是欧氏距离矩阵；表 10-12 所示是相似性分析的 Pearson 相关矩阵。相似性越强，相关系数越大，不相似性距离越小。例如，“月平均气温”与“月降雨量”相关系数最大，为 0.715；在不相似性的距离矩阵中，这两个变量间的距离最小，为 2.505。相反，在相似性测度的相关矩阵中，相关系数最小的是“月平均湿度”与“月平均日照时数”，为-0.051，它们的不相似性测度中的欧氏距离却是最大的，值为 4.808。

表 10-12 变量间的相似性测度，相关分析结果

	近似矩阵			
	值向量间的相关性			
	月平均气温(c)	月降雨量(mm)	月平均日照时数	月平均湿度
月平均气温(c)	1.000	.715	.690	.292
月降雨量(mm)	.715	1.000	.702	.384
月平均日照时数	.690	.702	1.000	-.051
月平均湿度	.292	.384	-.051	1.000

这是一个相似性矩阵

习 题 10

1. 什么是两个变量间的线性相关？两个变量间的相关系数的数值范围是什么？负相关系数反映的是两个变量数值间什么样的关系？
2. SPSS 提供了几个求相关系数的方法？各适合分析什么样的变量？
3. 数据文件 data10-05 中记录了 29 个被试者的身高、体重、肺活量的数据，试分析肺活量与哪个因素线性相关程度更高。说明为什么要计算偏相关。
4. 数据文件 data10-02 中是 474 名职工的职务等级 jobcat、起始工资 salbegin、当前工资 salary、受教育程度 educ、本单位工作经历(月)jobtime、以前工作经历(月)prevexp 的数据，id 为职工编号。分析该公司起始工资的确定与什么因素有关，当前工资与什么因素有关。
5. 数据文件 data10-06 是某公司太阳镜销售情况的数据。分析销售量与平均价格、广告费用和日照时间之间的关系。作图协助分析。本题使用偏相关分析是否有实际意义？

# 第11章 回归分析

回归分析(regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。它已广泛地应用于自然科学、社会科学等各个领域。按照回归分析中涉及的自变量的多少,可将回归分析分为一元回归分析和多元回归分析;按照自变量和因变量之间的关系类型,可将回归分析分为线性回归分析和非线性回归分析。在回归分析中,如果只包括一个自变量和一个因变量,且二者的关系可用一条直线近似表示,则称这种回归分析为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量,且因变量和自变量之间是线性关系,则称这种回归分析为多元线性回归分析。

在图 11-1 所示的【分析】菜单的【回归】子菜单中,对应的回归分析过程有以下几种:自动线性建模、线性回归(Linear)、曲线估计(Curve Estimation)、部分最小二乘回归(应为偏最小二乘回归)(Partial Least Squares Regression)、二元 Logistic 回归(二分变量 Logistic 回归)(Binary Logistic)、多项 Logistic 回归(多分变量 Logistic 回归)(Multinomial Logistic)、有序回归(定序回归)(Ordinal)、Probit(概率单位回归)、非线性回归(Nonlinear)、权重估计(加权估计)(Weight Estimation)、两阶最小二乘法(2-Stage Least Squares)、最优编码尺度回归(Optimal Scaling)。



图 11-1 【分析】菜单的【回归】子菜单

## 11.1 线性回归

自变量与因变量之间呈线形关系时,可以构造线性回归方程。线性回归包括一元线性回归和多元线性回归。

### 11.1.1 一元线性回归

#### 1. 一元线性回归方程

只有一个自变量的线性回归,称为一元线性回归,又称为直线回归。其分析的任务就是根据若干对观测 $(x_i, y_i)$  ( $i=1, 2, \dots, n$ )找出描述两个变量 $x$ 与 $y$ 之间关系的线性回归模型 $Y = \beta_0 + \beta_1 x + \varepsilon$ 。其中, $\varepsilon$ 是随机误差。求最优线性回归方程 $y = \beta_0 + \beta_1 x$ ,常用的方法是最小二乘法,也就是使该直线与各点的纵向垂直距离最小,即实测值 $y$ 与预测值 $\hat{y}$ 之差的平方和 $\sum (y - \hat{y})^2$ 达到最小。 $\sum (y - \hat{y})^2$ 也称为剩余(残差)平方和。因此,求回归方程 $y = \beta_0 + \beta_1 x$ 的

问题, 归根结底就是求  $\sum (y - \hat{y})^2$  取得最小值时  $\beta_0$  和  $\beta_1$  的问题。 $\beta_0$  称为截距,  $\beta_1$  为回归直线的斜率, 它们又共称为回归系数。

## 2. 一元线性回归方程的假设

德国数学家高斯提出 5 个假设, 满足这些假设的线性模型称为古典线性模型。

① 正态性假设: 随机误差项  $\varepsilon_i$  服从均值为 0, 方差为  $\sigma^2$  的正态分布。

② 等方差假设: 对所有  $x_i$ ,  $\varepsilon_i$  的条件方差同为  $\sigma^2$ , 且  $\sigma$  为常数, 即

$$\text{Var}(\varepsilon_i) = \sigma^2$$

③ 独立性假设即零均值假设: 在给定  $x_i$  的条件下,  $\varepsilon_i$  的条件期望值为 0, 即

$$E(\varepsilon_i) = 0$$

④ 无自相关性假设: 随机误差项  $\varepsilon$  的逐次观察值互不相关, 即

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j)$$

⑤  $\varepsilon$  与  $x$  的不相关性。假设随机误差项  $\varepsilon_i$  与相应的自变量  $x_i$  对因变量  $y$  的影响相互独立。换言之, 两者对因变量  $y$  的影响是可以区分的, 即  $\text{Cov}(\varepsilon_i, x_i) = 0$ 。

## 3. 一元线性回归方程的检验

为验证回归方程是否有统计学意义, 在根据原始数据求出回归方程后, 还需要对回归方程进行检验。检验的假设是总体回归系数为 0。可以选用下述 (1)~(3) 方法中的任意一种进行检验。此外, 还要对回归方程的预测效果进行检验。

(1) 回归系数的显著性检验。

① 对斜率检验的假设是总体回归系数  $\beta_1 = 0$ 。检验该假设的  $t$  值计算公式为

$$t = \frac{\hat{\beta}_1}{\text{SE}_b}$$

② 对截距检验的假设是总体回归方程截距  $\beta_0 = 0$ 。检验该假设的  $t$  值计算公式为

$$t = \frac{\hat{\beta}_0}{\text{SE}_a}$$

在以上两个公式中,  $\text{SE}_b$  是回归系数的标准误,  $\text{SE}_a$  是截距的标准误。

(2)  $R^2$  决定系数。它是判定线性回归直线拟合优度的重要指标, 公式为

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

它表明决定系数等于回归平方和在总平方和中所占的比率, 体现了回归模型所解释的因变量变异的百分比。 $R^2 = 0.775$ , 说明变量  $y$  的变异中有 77.5 % 是由变量  $x$  引起的,  $R^2 = 1$ , 表明因变量与自变量为函数关系;  $R^2 = 0$ , 表示自变量与因变量无线性关系。

(3) 方差分析。因变量观测值与均值之间差异的偏差平方和  $\text{SS}_t$  由两个部分组成, 表示为  $\text{SS}_t = \text{SS}_r + \text{SS}_e$ 。其中, 回归平方和  $\text{SS}_r$  反映了自变量  $x$  的重要程度; 残差平方和  $\text{SS}_e$  反映了试验误差以及其他意外因素对试验结果的影响。这两部分除以各自的自由度, 得到它们的均方

$$F = \frac{\text{回归均方}}{\text{残差均方}} = \frac{\sum (\hat{y} - \bar{y})^2 / p}{\sum (y - \hat{y})^2 / (n - p - 1)}$$

当  $F$  值太大时, 拒绝  $\beta_1 = 0$  的假设。

(4) Durbin-Watson 检验。在对回归模型的诊断中，需要诊断回归模型中误差项的独立性。如果误差项不独立，那么对回归模型的任何估计与假设所做出的结论都是不可靠的。

其参数称为  $DW$  或  $D$ ，取值范围是  $0 < D < 4$ ，统计学意义如下：

- ① 当残差与自变量互为独立时， $D \approx 2$ 。
- ② 当相邻两点的残差为正相关时， $D < 2$ 。
- ③ 当相邻两点的残差为负相关时， $D > 2$ 。

(5) 残差图示法。在直角坐标系中，常以预测值  $\hat{y}$  为横轴，以  $y$  与  $\hat{y}$  之间的误差  $e_i$  (或学生式残差值) 为纵轴，绘制残差的散点图。如果散点呈现明显的规律性，则认为存在自相关性，或者存在非线性、非常数方差的问题，见图 11-2(a)~(d)。

如果残差与因变量的关系类似图 11-2(a) 或 (b)，则需要对因变量或自变量进行变换；如果散点呈随机分布，则认为残差与因变量之间相互独立，见图 11-2(f)。

利用残差图还可以判断模型拟合效果。如果各点呈随机状，并绝大部分落在  $\pm 2\sigma$  范围 (68% 的点落在  $\pm \sigma$  之中，96% 的点落在  $\pm 2\sigma$  之中) 内，说明拟合效果较好，见图 11-2(f)；如果大部分点落在  $\pm 2\sigma$  范围之外，说明拟合效果不好，见图 11-2(g)。

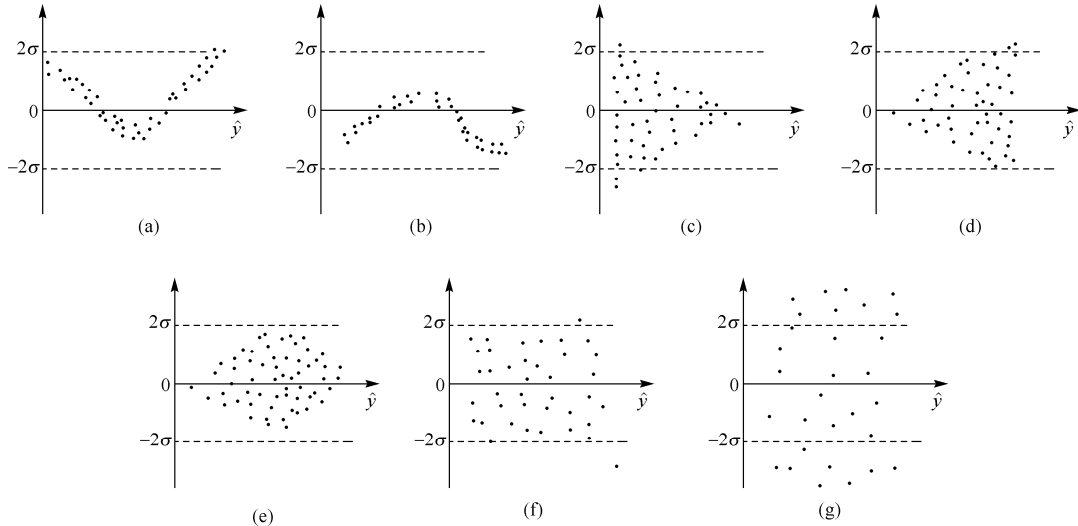


图 11-2 各种残差与预测值关系示意图

11.1.2 多元线性回归

1. 多元线性回归的概念

在实际问题中，影响因变量  $y$  的因素 (自变量，也称预报变量) 往往不止一个，而是多个，如  $x_1, x_2, \dots, x_p$ ， $p \geq 2$ 。在满足一定条件的前提下，可仿一元回归分析的做法，同样可以建立起描述多个自变量与一个因变量之间线性关系的多元线性回归模型。

多元回归模型为：

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon, \quad E(\varepsilon) = 0, \quad D(\varepsilon) = \sigma^2$$

式中， $\varepsilon$  是不可观测的随机误差， $E(\varepsilon)$  是其数学期望， $D(\varepsilon)$  是其方差，未知参数  $\beta_1, \beta_2, \dots, \beta_p$  称为回归系数， $\beta_0$  为回归常数， $x_1, x_2, \dots, x_p$  称为回归因子或预报因子。该模型除需要满足误差项  $\varepsilon_i$  之间相互独立且服从  $n$  维正态分布外，同样还必须满足 11.1.1 节中所述的各种假设。



通过实测  $n$  组独立的观测值  $(y_k; x_{k1}, x_{k2}, \dots, x_{kp})$ ,  $k=1, 2, \dots, n$ , 用最小二乘法可得到参数  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  的无偏估计  $\hat{\beta}_i (i=1, 2, \dots, p)$ , 通常还记作  $b_0, b_1, b_2, \dots, b_p$ 。这样可得到拟合后的经验回归方程  $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$ , 其中  $\hat{y}$  为根据所有自变量  $x$  计算出的估计值。为区别一元回归, 称  $b_1, b_2, \dots, b_p$  为  $y$  对应于  $x_1, x_2, \dots, x_p$  的偏回归系数。偏回归系数表示在其他所有自变量不变的情况下, 某一个自变量变化引起因变量变化的比率。

习惯上, 将这种根据多个自变量的最优组合建立回归方程来预测因变量的回归分析方法称为多元回归分析。

## 2. 多元线性回归分析中的统计指标

(1) 复相关系数  $R$  表示自变量  $x_i$  与因变量  $y$  之间线性关系密切程度的指标, 取值范围在  $0 \sim 1$  之间。其值越接近 1, 表示线性关系越强; 越接近 0, 表示线性关系越差。

(2)  $R^2$  判定系数与校正  $R^2$  判定系数。在多元回归中也使用  $R^2$  判定系数解释回归模型中自变量的变异在因变量变异中所占的比率。但是, 在多元回归中判定系数的值会随着进入回归方程的自变量的个数  $n$  或样本容量的大小的增加而增大。为了消除自变量的个数以及样本量的大小对判定系数的影响, 引进了校正  $R^2$  (Adjusted R Square)。校正  $R^2$  判定系数的公式为

$$\text{Adjusted } R^2 = 1 - \frac{\sum (y - \hat{y})^2 / (n - k - 1)}{\sum (y - \bar{y})^2 / (n - 1)}$$

式中,  $k$  为自变量的个数;  $n$  为观测数目。可以看出, 自变量数大于 1 时, 其值小于  $R^2$  判定系数。自变量数越多, 与  $R^2$  判定系数的差值越大。

(3) 零阶相关系数、部分相关系数与偏相关系数。

① 零阶相关系数 (Zero-Order)。各自变量与因变量之间的简单相关系数。

② 部分相关 (Part Correlation)。在排除了其他自变量对  $x_i$  的影响后, 当一个自变量进入回归方程模型后, 复相关系数的平方的增加量。

③ 偏相关系数 (Partial Correlation)。在排除了其他变量的影响后, 自变量  $x_i$  与因变量  $y$  之间的相关程度。部分相关系数小于偏相关系数。偏相关系数也可以用来作为筛选自变量的指标, 即通过比较偏相关系数的大小, 判别哪些变量对因变量具有较大的影响力。

## 3. 多元线性回归分析的检验

可以利用残差分析, 检验建立的回归模型是否很好地拟合了原始数据。还可以对回归方程中各自变量的系数进行检验, 以便在回归方程中保留那些有效影响因变量  $y$  值的自变量。

(1) 方差分析是对整个回归方程的显著性检验。检验的假设为: 总体的回归系数均为 0。使用统计量  $F$  进行检验, 其原理与一元回归的方程分析原理相同。

(2) 偏回归系数与常数项的检验。检验的假设为: 总体中回归方程各自变量偏回归系数为 0, 常数项为 0。检验使用  $t$  统计量。偏回归系数和常数项的  $T$  检验的公式分别为

$$t = \frac{\text{偏回归系数}}{\text{偏回归系数的标准误}}, \quad t = \frac{\text{常数项}}{\text{常数项的标准误}}$$

(3) 方差齐性检验。检验方差齐性是指残差的分布是常数, 与自变量或因变量无关。一般用绘制因变量预测值与学生式残差的散点图来检验。残差应随机地分布在一条穿过零点的水平直线的两侧。

(4) 残差的正态性检验。希望残差完全服从于正态分布也是不现实的,即使存在很理想的总体数据,其样本的残差的分布也只能是近似于正态分布。

在残差的正态性检验中,最直观、最简单的方法是作残差的直方图和累积概率图。

累积概率图(P-P 图)是用来判断一个变量的分布是否与一个指定的分布一致。如果两种分布基本相同,那么在 P-P 图中的点应该围绕在一条斜线的周围。通过观察残差(曲线)在假设直线(正态分布)周围的分布,可以判断是否符合正态分布。

### 11.1.3 异常值、影响点、共线性诊断

#### 1. 异常值的查找

异常值是指标准化残差过大的观测,在 SPSS 软件中,默认的判定标准是标准化残差的绝对值大于 3。

#### 2. 影响点的查找

因为影响点对参数估计的结果有较大的影响,所以要仔细地考虑在模型拟合时是保留还是剔除影响点。要注意,影响点的非标准化残差并不太大,因此需要仔细研判。

识别影响点的有效方法,是比较一个观测存在于回归方程时与不存在于回归方程中时残差的变化。主要的指标有:标准化残差、非标准化残差、学生氏剔除残差、学生化残差、剔除残差、Mahalanobis 距离、中心点杠杆值、COOK 距离、协方差比。

需要使用几个指标综合判断某一观测是否为影响点,使用个别指标可能会错判。

(1) 判别影响点的指标。

① 剔除残差(Dresid)。排除一个被认为是影响点的观测,回归分析的残差值。

② 学生化残差(Sdresid)。残差除以它的标准误,其值大于 2 时,应予以重视。

③ COOK 距离。它是对当一个被认为是影响点的观测被删除后,其他所有观测残差的变化量的测度。此值越大,表示这个被认为是影响点的观测的影响力越大。

④ Mahalanobis 距离。测定某一自变量观测与同一自变量所有观测平均值差异的统计量。此值越大,说明该观测为影响点的可能性越大。

⑤ 中心点杠杆值(Leverage Values)。当回归方程含有一个以上的自变量时,用来检测影响点的标准。其值在  $0 \sim (N-1)/N$  之间变化,杠杆值为 0 时,说明此观测值对回归方程没有影响;杠杆值接近  $(N-1)/N$ ,说明此观测对回归方程的贡献很大。从理论上说,希望数据所有的观测的杠杆值都接近于中心点杠杆平均值  $P/N$  ( $P$  为自变量数目),当杠杆值大于  $2P/N$  时,说明此观测的影响力很大。

⑥ 协方差比(Covariance Ratio)。它用来衡量某个观测是否对回归系数有显著的影响。当协方差比的值接近 1 时,表明此点的观测不是影响点。

国外有学者建议,当  $|\text{协方差比}-1| \geq 3P/N$  时,这个观测可以被视为影响点。

(2) 利用回归系数的变化检验影响点。Belsley 给出的建议是:仔细检查某一观测在与不在模型中前后变化的标准化  $\beta$  值,如果大于  $2/\sqrt{N}$  ( $N$  为观测的数目),那么此观测就有可能是影响点。

(3) 利用预测值来检测影响点。如果从模型中删除某一个观测后,其标准化预测值大于  $2/\sqrt{P/N}$  ( $P$  为自变量的个数,  $N$  为观测数)时,此观测有可能是影响点。

#### 3. 共线性问题

在回归方程中,各自变量对因变量虽然都是有意义的,但某些自变量彼此相关,就会存在

共线性的问题。这给评价自变量的贡献率带来困难。因此，需要对回归方程中的变量进行共线性诊断，并且确定它们对参数估计的影响。

共线性分为精确共线性与近似共线性。如果存在一些常数  $c_0, c_1, c_2$ ，使得等式  $c_1x_1 + c_2x_2 = c_0$  对数据中所有的观测都成立，则两个自变量  $x_1$  与  $x_2$  之间的关系为精确共线性；如果这个等式近似成立，那么两个自变量  $x_1$  与  $x_2$  之间的关系为近似共线性。

在只有两个自变量的情况下， $x_1$  与  $x_2$  共线性体现在两自变量间相关系数  $r_{12}$  上。精确共线性时  $r_{12}^2 = 1$ ；当它们之间不存在共线性时， $r_{12}^2 = 0$ 。 $r_{12}^2$  越接近于 1，共线性越强。

当自变量多于两个时， $x_i$  与其他自变量  $x$  之间的复相关系数的平方体现共线性，称为  $R_i^2$ 。它的值越接近 1，说明自变量之间的共线性程度越大。

当一组自变量精确共线性时，必须删除引起共线性的一个和多个自变量，否则不存在系数唯一的最小二乘估计。因为删除的自变量并不包含任何多余的信息，所以得出的回归方程并没有失去什么。当为近似共线性时，一般将引起共线性的自变量删除，但需要掌握的原则是：务必使丢失的信息最少。识别共线性的统计量有以下几个：

① 容忍度 (Tolerance)。定义为  $\text{Tol}_i = 1 - R_i^2$ ，其值介于 0~1 之间。其值越小，自变量  $x_i$  与其他自变量  $x$  之间的共线性越强。使用容忍度作为共线性量度指标的条件比较严格，观测一定要近似于正态分布。

② 方差膨胀因子 (VIF)。定义为  $\text{VIF}_i = 1/(1 - R_i^2)$ ，是容忍度的倒数，其值介于 1~ $\infty$  之间。其值越大，自变量之间存在共线性的可能性越大。

有专家认为，容忍度小于 0.1 或 0.2，或者 VIF 值大于 5 或 10 可以认为存在共线性问题，读者可以参考。

③ 特征值 (Eigenvalues)。当若干特征值较小并且接近 0 时，说明某些变量之间存在很高的相关性。这些变量的观测出现较小的变化时，会导致回归系数较大的变化。

④ 条件指数 (Condition Index)。是在计算特征值时产生的一个统计量。其值越大，说明自变量间的共线性的可能性越大。一般认为，条件指数  $\geq 15$  时可能存在共线性问题；条件指数  $\geq 30$  时存在严重的共线性问题。 $\text{Condition Index} = \sqrt{\text{最大特征值} / \text{第} i \text{个特征值}}$ 。

⑤ 方差比例 (Variance Proportions)。同一序号的特征值对应的变量的方差比例。比例越大，其共线性的可能性越大。

⑥ 常用的共线性问题的解决方法：

- 从产生共线性问题的自变量中剔除不重要的自变量。
- 增加样本量。
- 重新抽取样本数据。不同样本的观测的共线性是不一致的，所以重新抽取样本数据有可能减少共线性问题的严重程度。
- 另外，采用主成份回归、岭回归、偏最小二乘法、LASSO 回归等也可以解决多重共线性问题。

#### 11.1.4 变非线性关系为线性关系

因变量与自变量的关系不是线性关系，但利用其他方法也未能很好地拟合数据时，就需要进行数据的非线性到线性关系的转换。如果因变量或残差不符合假设条件，也需要进行转换。非线性转换为线性关系的原则及方法的统计学知识已经超出本书范围，读者可以参考有关书籍，在此仅给予提示。

- (1) 当残差的分布呈现正偏态时,对因变量进行对数转换。当残差的分布呈现负偏态分布时,采用平方根转换。
- (2) 如果残差的方差呈现不稳定状态,可用表 11-1 的方法进行校正,注意适用条件。

表 11-1 变量转换公式表

转换方法	使用条件	注 释
$\sqrt{y}$	$\text{Var}(e_i) \propto E(y_i)$	因变量服从泊松分布
$\sqrt{y} + \sqrt{y+1}$	$\text{Var}(e_i) \propto E(y_i)$	某些因变量的值为 0 或者很小
$\lg y$	$\text{Var}(e_i) \propto [E(y_i)]^2, y > 0$	因变量的值的范围很大
$\lg(y+1)$	$\text{Var}(e_i) \propto [E(y_i)]^2$	因变量的某些值为 0
$1/y$	$\text{Var}(e_i) \propto [E(y_i)]^4$	因变量的值集中在 0 的附近,当自变量明显降低时,因变量出现较大的值。例如,自变量是治疗某病的药剂量,因变量是反应时间
$1/(y+1)$	$\text{Var}(e_i) \propto [E(y_i)]^4$	某些自变量为 0 的情况
$\arcsin \sqrt{y}$	$\text{Var}(e_i) \propto E(y_i)(1 - (y_i))$	用于二项比例 ( $0 \leq \text{因变量} \leq 1$ )

注:  $\text{Var}(e_i)$  为  $e_i$  的方差;  $e_i$  为第  $i$  个观测的统计误差;  $E(y_i)$  为随机变量  $y_i$  的算术平均数。

当方差随着因变量的增大或减小而变化时,  $\sqrt{y}$ 、 $\lg(y)$  与  $1/y$  都是可以选用的方法,但是它们转换的力度却是依次递增的;当因变量是直到某一事件发生和完成的时间,则常使用倒数或逆变换;当因变量的数据中出现 0 或负数时,为了避免对数或开根号没有意义的情况出现,常采用  $(y+\text{常数})$  的方法,常数一般取 1;在经济学研究方面,  $\lg(y)$  是一种常用的方法。

(3) 非线性数据转变为线性数据的方法主要包括:取对数、取倒数和取平方根。注意,并非所有的函数都是可以线性化的。

- ① 当回归方程有可能是多项式方程,如  $y=x^2+3x+1$  时,可以取平方根或取倒数。
- ② 当要建立的回归方程未知时,可以利用散点图发现规律,进行转换,见表 11-2。

表 11-2 转换为线性的常用方法

变化方法		回 归 式
$\lg y$	$\lg x$	$y = \alpha x$
$\lg y$	$x$	$y = \alpha e^{\beta x}$
$y$	$\lg x$	$y = \alpha + \beta \log x$
$1/y$	$1/x$	$y = x/(\alpha + \beta)$
$1/y$	$x$	$y = 1/(\alpha + \beta x)$
$y$	$1/x$	$y = \alpha + \beta(1/x)$

11.1.5 线性回归过程

1. 数据要求

(1) 自变量与因变量应该是数值型变量,类似研究领域、居住地区、信仰等分类变量应重新编码为哑变量或者其他类型的对比变量。

(2) 假设。对自变量的每一个值,因变量的分布必须是正态的。因变量方差的分布对所有自变量的值都应该是一个常数。因变量和每个自变量之间的关系应该是线性的,所有观测应该是独立的。

在进行回归分析之前,最好用图形探索因变量随自变量变化的趋势,以便确定数据是否适合线性模型。通过散点图还可以发现异常值。

2. 建立线性模型的操作步骤

(1) 按【分析→回归→线性】顺序打开如图 11-3 所示的【线性回归】主对话框。



图 11-3 【线性回归】主对话框

(2) 在源变量框中选择一个因变量进入【因变量】框, 选择一个或多个自变量进入【自变量】框。

可以利用【上一张(前一个模型)】与【下一张(下一个模型)】按钮切换, 选择不同的自变量组构建不同的模型; 每个模型中可以对不同自变量组采用不同的分析方法, 如有的自变量组采用【进入(强行进入)】法, 有的采用【逐步(逐步筛选)】法。构建的模型按顺序保存第  $n$  个模型中。

(3) 在【方法】框中选择所需的回归分析中自变量进入回归方程的方法。

①【进入(强行进入)】法。所选择的自变量全部进入回归模型。这是默认方式。

②【逐步(逐步筛选)】法。根据在【线性回归: 选项】对话框中所设定的判定标准, 选择符合判定标准的且对因变量贡献最大的自变量进入回归方程。然后将模型中符合剔除判定标准的变量移出模型, 重复进行直到回归方程中的自变量均符合进入模型的判定标准, 而模型外的自变量都不符合进入模型的判定标准为止。

③【删除(剔除)】法。先建立全模型, 再按剔除标准剔除自变量。

④【向前】选择法。从模型中无自变量开始, 根据在【线性回归: 选项】对话框中所设定的判定标准, 每次将一个最符合判定标准的变量引入模型, 直至所有符合判定标准的变量都进入模型为止。第一个引入回归模型的变量应该是与因变量的相关系数绝对值最大的变量。如果指定的判定标准是  $F$  值, 每次将方差分析的  $F$  值最大且大于指定的  $F$  值的变量引入模型。如果指定的判定标准是大于  $F$  值的概率, 每次将概率最小且小于指定的概率的变量引入模型。

⑤【向后】剔除法。先建立全变量模型。模型中与因变量具有最小偏相关的变量若符合在【线性回归: 选项】对话框中所设定的判定标准, 被最先从模型中剔出, 然后根据设定的判定标准, 重复以上步骤, 直到回归方程中不再含有符合剔出判定标准的自变量为止。

(4) 根据一个设定的变量值, 选择参与回归分析的观测。将选择变量送入【选择变量】框中, 单击【规则】按钮, 打开如图 11-4 所示的对话框。

在下拉列表中选择关系运算法则: 等于、不等于、小于、小于等于、大于、大于等于。然后在【值】框中输入判定标准, 最后单击【继续】按钮。

(5) 在主对话框中, 选择一个变量进入【个案标签】框, 其值作为观测标签。

(6) 选择一个作为权重的变量进入【WLS 权重】框中。利用加权最小平方方法给观测不同的权重值, 它可用来补偿或减少采用不同测量方式所产生的误差。

因变量与自变量, 不能再作为加权变量使用, 加权变量的值如果为零、负数或缺失值, 那么相对应的观测将被删除。

(7) 单击【统计量】按钮, 打开如图 11-5 所示的对话框, 选择要输出的统计量。



图 11-4 【线性回归: 设置规则】对话框



图 11-5 【线性回归: 统计量】对话框

① **【回归系数】** 栏。有关回归系数的选项。

- **【估计】**。输出回归系数  $B$ 、 $B$  的标准误、标准化回归系数 Beta、对回归系数为 0 的假设进行检验的  $T$  值,  $T$  值的双侧检验的显著性概率 Sig。
- **【置信区间】**。输出每一个非标准化回归系数 95% 的置信区间或者一个方差矩阵。
- **【协方差矩阵】**。输出非标准化回归系数的协方差矩阵、各变量的相关系数矩阵。

② 与模型拟合及其拟合效果有关的选项。

- **【模型拟合度】**。对拟合过程中引入模型及从模型中剔除的变量, 输出复相关系数  $R$ 、其平方  $R^2$ , 及其修正值, 估计值的标准误, ANOVA 方差分析表。这是默认选项。
- **【 $R$  方变化】**。输出  $R^2_{\text{ch}}$ 、 $F_{\text{ch}}$ 、 $\text{Sig}_{\text{ch}}$ 。 $R^2_{\text{ch}}$  是当回归方程引入或剔除一个自变量后  $R^2$  统计量的变化量。如果与某个自变量有关的  $R^2$  变化较大, 说明进入和从回归方程剔除的可能是一个较好的回归自变量。
- **【描述性】**。输出有效观测的数量、变量的平均数、标准差、相关系数矩阵及其单侧检验显著性水平矩阵。
- **【部分相关和偏相关系数】**。输出部分相关系数、偏相关系数与零阶相关系数。
- **【共线性诊断】**。输出用来诊断各变量共线性问题的各种统计量和容限值。

③ **【残差】** 栏。有关残差分析的选项。

- **【Durbin-Watson】**。输出 Durbin-Watson 统计量以及可能是异常值的观测诊断表。
- **【个案诊断】**。输出观测诊断表。
- **【离群值】**。设置异常值的判定标准, 默认值为  $\geq 3$ 。
- **【所有个案】**。输出所有观测的残差值。

(8) 单击 **【绘制】** 按钮, 打开如图 11-6 所示的对话框。选择要输出的图形。默认情况下不输出图形。

① 在左侧的源变量框中, 根据需要选择两个变量的组合, 并分别送入 X、Y 轴变量框中。

可以选择的作图元素有: DEPENDENT(因变量)、ZPRED(标准化预测值)、ZRESID(标准化残差)、DRESID(剔除残差)、ADJPRED(修正后预测值)、SRESID(学生化残差)、SDRESID(学生氏剔除残差)。

② **【标准化残差图】** 栏。输出标准化残差图。

- **【直方图】**。输出带有正态曲线的标准化残差的直方图。
- **【正态概率图】**。输出 P-P 图, 即残差的正态概率图, 检查残差的正态性。

③ **【产生所有部分图】**。输出每一个自变量的残差相对于因变量残差的散布图。

(9) 单击 **【保存】** 按钮, 打开如图 11-7 所示的 **【线性回归: 保存】** 对话框, 指定要保存到数据窗口的新变量。

① **【预测值】** 栏。可选择输出的预测值有: **【未标准化】** 预测值、**【标准化】** 预测值、**【调节(应为修正值)】** (将一个观测值排除在回归方程之外时, 它本身的预测值)、**【均值预测值的 S.E.】**。

② **【距离】** 栏。选择要输出的距离。选项有: **【Mahalanobis 距离】**、**【Cook 距离】**、**【杠杆值】** (中心点杠杆值)。

③ **【预测区间】** 栏。选择输出预测区间可选项有:

- **【均值】**。预测区间高低限的平均值。
- **【单值】**。观测预测值上、下限的间距。

选择上述两项, 要在 **【置信区间】** 框中指定可信区间, 默认为 95%, 可输入 0~99.99 之间的值。



图 11-6 【线性回归：图】对话框

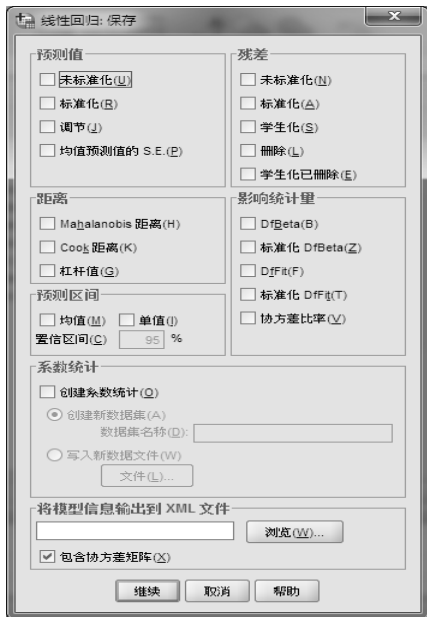


图 11-7 【线性回归：保存】对话框

④【残差】栏。选择输出的残差有：【未标准化】残差、【标准化】残差、【学生化】残差、【删除(应为剔除)】残差、【学生化已删除(应为学生化剔除)】残差。

⑤【影响统计量】栏。输出影响点的统计量。

- 【DfBeta】。因排除一个特定的观测值所引起的回归系数的变化值。一般情况下，如果此值大于界值 $|2/\sqrt{N}|$ ，则被排除的观测值有可能是影响点。
- 【标准化 DfBeta】值。在数据文件中保存的变量名为 SDBM\_N， $M=0$  时为常数项， $M \geq 1$  时为自变量； $N \geq 1$ ，为  $N$  次运行的模型编号。
- 【DfFit】。因排除一个特定的观测值所引起的预测值的变化量。
- 【标准化 DfFit】。标准化 DfFit 值大于界值 $|2\sqrt{P/N}|$ 的观测值时可认为是影响点。
- 【协方差比率矩阵】。剔除了一个影响点的协方差矩阵与全部观测的协方差矩阵的比。比值接近 1，说明观测对方差矩阵没有显著影响。

⑥【系数统计】栏。在同一会话中可以继续使用数据集，除非在会话结束之前明确将其保存为文件，否则不会将其另存为文件。选择该栏中的【创建系数统计】选项，可将回归系数保存到数据集或数据文件。

- 【创建新数据集】。可在【数据集名称】框中输入数据集名称。数据集名称必须符合变量命名规则。
- 【写入新数据文件】。选择本选项后单击【文件】按钮，则在弹出的输出【文件】选项卡中，将回归系数保存到一个指定的文件中。

⑦【将模型信息输出到 XML 文件】框。可将模型的信息即参数估计值及其(可选)协方差输出到指定的 XML 格式的文件中，以备他用。单击【浏览】按钮可指定保存位置和文件名。

(10) 单击【选项】按钮，打开如图 11-8 所示的【线性回归：选项】对话框。



图 11-8 【线性回归：选项】对话框

- ① **【步进方法标准】** 栏。设置变量引入模型或从模型剔除的判定标准。
- **【使用 F(检验) 的概率】**。采用 F 检验的概率值作为判定标准。一个自变量 F 检验 Sig. 值 $\leq$ 进入值时, 该变量被引入回归方程中; 当一个变量 F 检验的 Sig. 值 $\geq$ 剔除值时, 该变量从回归方程中剔除。系统默认 **【进入】** 值为 0.05, **【删除(即剔除)】** 值为 0.10。用户可以自定义这两个值, 但必须满足 **删除值** $>$ **进入值 $>$ 0。加大**进入值**可使更多变量能够进入方程, 减小**删除值**可从方程中剔除更多的变量。**
  - **【使用 F 值】**。采用 F 值作为判定标准。系统默认 F 值 $\geq 3.84$  的变量被选入模型中; F 值 $\leq 2.71$  的变量从模型中剔除。可以自定义 **【进入】** 值、**【删除】** 值, 但必须满足 **删除值** $>$ **进入值 $>$ 0。减少**进入值**可能使更多的变量能够进入方程; 加大**删除值**可能剔除更多的变量。**
- ② **【在等式中包含常量】**。在回归方程中包含常数项。这是默认选项。
- ③ **【缺失值】** 栏。进行缺失值处理。
- **【按列表排除个案(应为按列排除个案)】**。将变量表中变量具有缺失值的所有观测排除在计算之外。
  - **【按对排除个案】**。剔除计算相关系数的一对变量中含有缺失值的观测。
  - **【使用均值替换】**。利用变量的平均数代替缺失值。

11.1.6 线性回归分析实例

**【例 1】** 使用数据文件 data11-01, 建立一个以 salbegin(初始工资)、prevexp(工作经验)、educ(受教育年数)为自变量, salary(当前工资)为因变量的回归模型。

1) 作数据散点图

观察因变量与自变量之间关系是否有线性特点。

(1) 按 **【图形→旧对话框→散点图/点图→简单分布→定义】** 顺序打开 **【简单散点图作图】** 对话框。

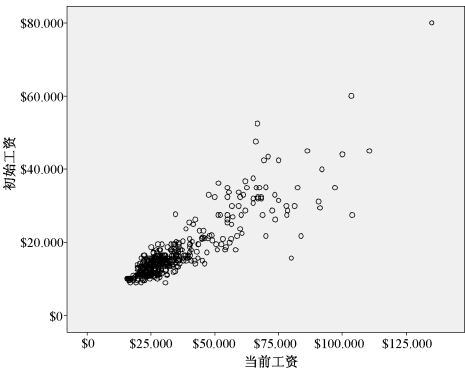


图 11-9 初始工资与当前工资散点图

(2) 将变量 salbegin、salary 依次选作 Y 轴变量与 X 轴变量, 单击 **【确定】** 按钮。

在输出窗中生成的图形见图 11-9, 其 Y 轴为初始工资, X 轴为当前工资。根据同样操作方法可以作以 salary 为 Y 轴, 分别以其他几个自变量为 X 轴的散点图。

从图 11-9 中看出, 初始工资与当前工资存在明显的线性关系, 以初始工资为自变量建立线性回归方程是可能的。对其他可能引入模型的变量, 也应该作出散点图, 有助判断。应当注意, 在最终确定回归方程结果之前还应审查数据中的奇异值、影响点。另外, 两个变量作对数变换后, 线性关系会更好。读者可以自己作图验证。

2) 回归模型的建立

(1) 按 **【分析→回归→线性】** 顺序打开线性模型主对话框。

(2) 在左侧的源变量框中选择变量 salary 作为因变量进入 **【因变量】** 框中。选择变量 salbegin、prevexp、jobtime、educ 作为自变量进入 **【自变量】** 框中。



(3) 在【方法】框中选择【逐步】回归法。

(4) 单击【统计量】按钮，打开如图 11-5 所示的对话框。选择【估计】和【模型拟合度】选项，输出各种常用统计量；在【残差】栏中选择【个案诊断】项，要求进行奇异值判别，并在【离群值】框中输入“3”，设置观测标准差大于等于 3 为奇异值；单击【继续】按钮返回。

(5) 单击【保存】按钮，打开如图 11-7 的对话框。选择【Mahalanobis 距离】、【Cook 距离】、【杠杆值】、【标准化 Dfbeta】、【标准化 DfFit】和【协方差比率】选项，这些统计量将保存在数据文件中，用来确定影响点，单击【继续】按钮返回。

(6) 为了检测模型的直线性和方差的齐性，作散点图。单击【绘制】按钮打开图 11-6 所示【制图】对话框，将变量 ZPRED 与 ZRESID 分别选入【X】、【Y】框中。单击【继续】按钮返回主对话框。

(7) 单击【确定】按钮，提交系统执行。

3) 输出结果(见表 11-3~表 11-9 及图 11-10~图 11-12)

表 11-3 引入或从模型中剔除的变量

输入/移去的变量 <sup>a</sup>			
模型	输入的变量	移去的变量	方法
1	起始工资		步进(准则: F-to-enter 的概率 <= .050, F-to-remove 的概率 >= .100)。
2	过去经验(月)		步进(准则: F-to-enter 的概率 <= .050, F-to-remove 的概率 >= .100)。
3	受雇月数		步进(准则: F-to-enter 的概率 <= .050, F-to-remove 的概率 >= .100)。
4	受教育程度(年)		步进(准则: F-to-enter 的概率 <= .050, F-to-remove 的概率 >= .100)。

a. 因变量: 当前工资

表 11-4 拟合过程小结

模型汇总 <sup>e</sup>				
模型	R	R 方	调整 R 方	标准估计的误差
1	.880 <sup>a</sup>	.775	.774	\$8,115.356
2	.891 <sup>b</sup>	.793	.793	\$7,776.652
3	.897 <sup>c</sup>	.804	.803	\$7,586.187
4	.900 <sup>d</sup>	.810	.809	\$7,465.139

- a. 预测变量: (常量), 起始工资。
- b. 预测变量: (常量), 起始工资, 过去经验(月)。
- c. 预测变量: (常量), 起始工资, 过去经验(月), 受雇月数。
- d. 预测变量: (常量), 起始工资, 过去经验(月), 受雇月数, 受教育程度(年)。
- e. 因变量: 当前工资

表 11-5 方差分析

Anova <sup>a</sup>					
模型	平方和	df	均方	F	Sig.
1 回归	1.068E+11	1	1.068E+11	1622.118	.000 <sup>b</sup>
残差	31085446686	472	65858997.22		
总计	1.379E+11	473			
2 回归	1.094E+11	2	54716073578	904.752	.000 <sup>c</sup>
残差	28484348280	471	60476323.31		
总计	1.379E+11	473			
3 回归	1.109E+11	3	36955960955	642.151	.000 <sup>d</sup>
残差	27048612571	470	57550239.51		
总计	1.379E+11	473			
4 回归	1.118E+11	4	27944979881	501.450	.000 <sup>e</sup>
残差	26136575912	469	55728306.85		
总计	1.379E+11	473			

- a. 因变量: 当前工资
- b. 预测变量: (常量), 起始工资。
- c. 预测变量: (常量), 起始工资, 过去经验(月)。
- d. 预测变量: (常量), 起始工资, 过去经验(月), 受雇月数。
- e. 预测变量: (常量), 起始工资, 过去经验(月), 受雇月数, 受教育程度(年)。

表 11-6 逐步回归过程中不在方程中的变量

已排除的变量 <sup>a</sup>					
模型		Beta In	t	Sig.	共线性统计量
1	过去经验(月)	-.137 <sup>b</sup>	-6.558	.000	-.289 .998
	受雇月数	.102 <sup>b</sup>	4.750	.000	.214 1.000
	受教育程度(年)	.172 <sup>b</sup>	6.356	.000	.281 .599
2	受雇月数	.102 <sup>c</sup>	4.995	.000	.225 1.000
	受教育程度(年)	.124 <sup>c</sup>	4.363	.000	.197 .520
3	受教育程度(年)	.113 <sup>d</sup>	4.045	.000	.184 .516

- a. 因变量: 当前工资
- b. 模型中的预测变量: (常量), 起始工资。
- c. 模型中的预测变量: (常量), 起始工资, 过去经验(月)。
- d. 模型中的预测变量: (常量), 起始工资, 过去经验(月), 受雇月数。

表 11-3 自左至右各列含义分别为：拟合步骤编号、每步引入回归方程的自变量、从回归方程中被剔除的自变量、自变量引入或剔除出方程的判定标准。可以看出，4 个被选择的自变量经过逐步回归过程都进入了回归方程，没有变量被剔除。

表 11-4 自左至右各列含义分别为：回归方程模型编号、回归方程的复相关系数  $R$ 、 $R^2$ 、修正的  $R^2$ 、估计的标准误。一般随着模型中变量个数的增加， $R^2$  的值也在不断增加，而修正  $R^2$  值与变量的数目无关。本例这个特点不明显。 $R^2$  值的增加并不意味着模型更好，也未必会减少估计的标准误。修正  $R^2$  值能较确切地反映拟合优度，因此一般从修正  $R^2$  值看拟合优度。除非需要，自变量数量不应太多，多余的自变量会给解释回归方程造成困难。包含多余自变量的模型不但不会改善预测值，反而有可能增加标准误差。由表 11-4 的  $R^2$  以及修正的  $R^2$  值可以看出建立的回归方程比较好。

表 11-5 为方差分析表，显示了回归拟合过程中每一步的方差分析结果。Sig. 为  $F$  值大于  $F$  临界值的概率。方差分析结果表明，当回归方程包含不同的自变量时，其显著性概率值均小于 0.001，即拒绝回归系数均为 0 的原假设。因此，最终的回归方程应该包括这 4 个自变量，且方程拟合效果很好。

表 11-6 显示了每一步回归过程中不在方程中的变量信息。

第一步是方程中已经有了 1 个变量 `salbegin`，外面有 3 个变量。如果每个外面的变量单独进入模型，则形成两个自变量模型的统计量及检验结果，以及模型中两个自变量之间的共线性诊断。显然，与因变量 `salary` “当前工资” 相关绝对值最高的是“工作经验”。如果它进入模型， $T$  检验的显著性小于 0.001，拒绝回归系数为 0 的假设。共线性诊断容忍度接近 1，说明它与第一个进入模型的自变量不具共线性，所以自变量 `prevexp`(工作经验) 第二个进入模型。其他步的分析与此相同。

表 11-7 各步回归过程中的统计量

系数 <sup>a</sup>								
模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准误差	试用版			容差	VIF
1	(常量)	1928.206	888.680		2.170	.031		
	起始工资	1.909	.047	.880	40.276	.000	1.000	1.000
2	(常量)	3850.718	900.633		4.276	.000		
	起始工资	1.923	.045	.886	42.283	.000	.998	1.002
	过去经验(月)	-22.445	3.422	-.137	-6.558	.000	.998	1.002
3	(常量)	-10266.629	2959.838		-3.469	.001		
	起始工资	1.927	.044	.888	43.435	.000	.998	1.002
	过去经验(月)	-22.509	3.339	-.138	-6.742	.000	.998	1.002
	受雇月数	173.203	34.677	.102	4.995	.000	1.000	1.000
4	(常量)	-16149.671	3255.470		-4.961	.000		
	起始工资	1.768	.059	.815	30.111	.000	.551	1.814
	过去经验(月)	-17.303	3.528	-.106	-4.904	.000	.865	1.156
	受雇月数	161.486	34.246	.095	4.715	.000	.992	1.008
	受教育程度(年)	669.914	165.596	.113	4.045	.000	.516	1.937

a. 因变量: 当前工资

注意：表中给出的是中文变量标签。以下模型中的变量均为变量名。

表 11-7 给出了每一步回归过程的统计量及检验结果。

$B$  为非标准化回归系数，也称偏回归系数，它是在控制了其他变量之后得到的。只有当所有的自变量单位统一时，它们才有可比性。由方差分析得出回归方程有统计意义，而回归方程中的每一个偏回归系数不一定都有显著性，但至少要有一个是显著的。

标准系数 Beta(软件中翻译为“试用版”，有误)是标准化回归系数，具有可比性，是所有的变量按统一方法标准化后拟合的回归方程中各标准化变量的系数。

t 为偏回归系数为 0(和常数项为 0)的假设检验的 t 值，具有较好预测效果的变量的 t 值应大于 2 或者小于-2。Sig.为假设检验的显著性概率。4 步回归的各变量和常数项的检验的 p 值均小于 0.05。

共线性统计量给出了容忍度值(表中定义为容差)和方差膨胀因子值(VIF)。

以第二步为例说明这些统计量：回归方程中包含常数项(常量)和自变量 salbegin、prevexp，因变量为 salary。共线性诊断的指标：容忍度(Tolerance)分别为 0.998、0.998，接近 1，方差膨胀因子 VIF 值都不大，可以认为两个自变量之间不存在共线性问题。

模型 2: salary = 3850.7+1.92salbegin - 22.4prevexp。方程常数项和两个自变量 T 检验的显著水平值均小于 0.001，拒绝常数项和回归系数为 0 的假设。方程成立。

模型 5: salary = -16149.7+1.77salbegin - 17.30prevexp + 161.49jobtime + 669.9educ。是最后的回归模型。每个自变量的显著水平值都小于 0.001。各个自变量的容忍度值分别为 0.551、0.865、0.992、0.516，没有出现特别小的数值；相应的 VIF 值分别为 1.814、1.156、1.008、1.937，没有很大的数值出现，说明方程中各自变量之间没有出现共线性问题。

表 11-8 所示为异常值诊断表。自左至右各列的含义分别为：奇异值观测编号(表中译为“案例数目”，不准确)、标准化残差、因变量当前工资的值、当前工资的预测值、残差。表中给出了被怀疑为异常值的观测的编号，这些观测之所以被怀疑为异常值，是因为它们的标准化残差绝对值都大于设置值 3。

表 11-9、图 11-10 和图 11-11 配合可查找影响点。表 11-9 中的“Mahal·距离”、“Cook 的距离”、“居中杠杆值”统计量值，都可以帮助判断是否含有影响点。以“Mahal·距离”为例，其值范围越大，越可能含有影响点。

表 11-8 当前工资变量的异常值表

案例诊断 <sup>a</sup>				
案例数目	标准 残差	当前工资	预测值	残差
18	6.173	\$103,750	\$57,671.26	\$46,078.744
103	3.348	\$97,000	\$72,009.89	\$24,990.108
106	3.781	\$91,250	\$63,026.82	\$28,223.179
160	-3.194	\$66,000	\$89,843.83	-\$23,843.827
205	-3.965	\$66,750	\$96,350.44	-\$29,600.439
218	6.108	\$80,000	\$34,405.27	\$45,594.728
274	5.113	\$83,750	\$45,581.96	\$38,168.038
449	3.590	\$70,000	\$43,200.04	\$26,799.959
454	3.831	\$90,625	\$62,027.14	\$28,597.858

a. 因变量: 当前工资

表 11-9 残差统计量

残差统计量 <sup>a</sup>					
	极小值	极大值	均值	标准 偏差	N
预测值	\$13,354.82	\$150,076.77	\$34,419.57	\$15,372.742	474
标准 预测值	-1.370	7.524	.000	1.000	474
预测值的标准误差	391.071	3191.216	721.093	260.806	474
调整后的预测值	\$13,290.94	\$153,447.97	\$34,425.45	\$15,451.094	474
残差	-\$29,600.439	\$46,078.746	-\$0.000	\$7,433.507	474
标准 残差	-3.965	6.173	.000	.996	474
Student 化 残差	-4.089	6.209	.000	1.004	474
已删除的残差	-\$31,485.213	\$46,621.117	-\$5.882	\$7,553.608	474
Student 化 已删除的残差	-4.160	6.474	.002	1.016	474
Mahal·距离	.300	85.439	3.992	5.306	474
Cook 的距离	.000	.223	.003	.016	474
居中杠杆值	.001	.181	.008	.011	474

a. 因变量: 当前工资

	MAH_1	COO_1	COV_1	SDF_1
28	3.05908	.00005	1.01914	-.01635
29	85.43873	.22320	1.17231	-1.0609
30	3.02689	.00019	1.01821	-.03072
31	2.63989	.00070	1.01367	.05909
32	15.38601	.05906	.95831	.54764
33	2.84042	.00156	1.00871	.08818
34	11.30935	.01472	1.00761	.27178

图 11-10 判定影响点的各种常用统计量

	SDB0_1	SDB1_1	SDB2_1	SDB3_1	SDB4_1
29	.14326	-1.0054	.11259	-.22739	.48914
30	.02116	.01126	.00131	-.02152	-.01166
31	-.0250	.00466	-.0098	.04701	-.01787
32	-.2404	.41245	-.0205	.21563	-.07845
33	-.0631	-.02912	.00103	.06343	.03278
34	-.1561	.15770	.05034	.11722	.01547

图 11-11 标准化回归系数变化量

在图 11-10 中，根据前述的判定标准，可以大致断定 29、32 号观测为影响点。但是，进一步观察后发现 34 号观测也存在是影响点的可能性。为判别某一观测是否为影响点，可以比较此观测在与不在回归方程中时，标准化回归系数的差异，如图 11-11 所示。

图 11-11 所示是数据文件中的新变量。SDB0\_1~SDB4\_1 分别对应常数项和第 1~4 个自变量的标准化回归系数的变化量。34 号观测的 SDB0\_1、SDB1\_1、SDB3\_1 的值大于  $2/\sqrt{N}$  (本例  $N = 474$ ，值为 0.09186)，即常数项、第 1、3 个自变量 jobtime、educ 的标准化回归系数变异较大。以此初步认定 34 号观测也为影响点。当然对影响点的判断仅凭借一个指标往往是不充分的，还需要用其他指标进行比较判断。

图 11-12 所示为“当前工资”的标准化预测值与其学生化残差散点图，可以看到绝大部分观测随机地落在垂直围绕  $\pm 2$  的范围内，预测值与学生化残差值之间没有明显的关系，所以回归方程应该满足线性与方差齐性的假设且拟合效果较好。

【例 2】为说明共线性的诊断的方法，仍使用数据文件 data11-01，以其中的 salbegin、prevexp、jobtime、educ、age 作为自变量，salary 作为因变量，用强行进入法建立所有自变量都进入回归方程的全模型。

打开数据文件 data11-01，进入【线性回归】主对话框，选择变量 salary 作为因变量，选择 salbegin、prevexp、jobtime、educ、age 作为自变量。在【方法】框中选择【进入(强行进入)】方法。在【线性回归：统计量】对话框中选择【共线性诊断】，要求进行共线性诊断，单击【确定】按钮进行分析。结果见表 11-10。

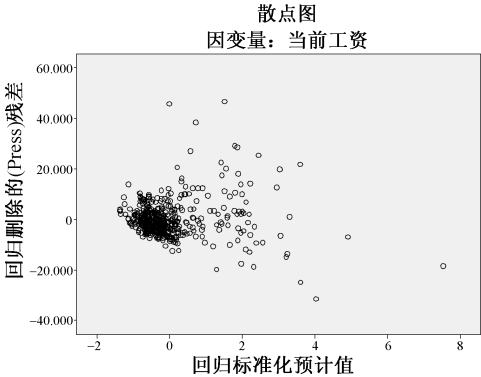


图 11-12 当前工资的预测值与学生化残差散点图

表 11-10 共线性诊断指标

共线性诊断<sup>a</sup>

模型	维数	特征值	条件索引	方差比例					
				(常量)	起始工资	受雇月数	过去经验(月)	受教育程度(年)	年龄
1	1	5.313	1.000	.00	.00	.00	.00	.00	.00
	2	.509	3.232	.00	.01	.00	.30	.00	.00
	3	.136	6.243	.01	.49	.01	.03	.00	.01
	4	.022	15.706	.00	.34	.00	.46	.51	.36
	5	.014	19.226	.00	.13	.49	.14	.32	.40
	6	.006	30.287	.99	.02	.49	.07	.17	.23

a. 因变量: 当前工资

(a)

系数<sup>a</sup>

模型		共线性统计量	
		容差	VIF
1	起始工资	.551	1.815
	受雇月数	.983	1.018
	过去经验(月)	.347	2.882
	受教育程度(年)	.508	1.967
	年龄	.347	2.880

a. 因变量: 当前工资

(b)

表 11-10(a) 中的“特征值”一列有 3 个特征值分别为 0.022、0.014、0.006，都非常接近 0，对应的 3 个条件指数(表中译为“条件索引”，不准确)分别为 15.706、19.226、30.287 都大于 15，

两个指标都说明在这 3 个变量间可能存在共线性。条件指数大于 30 则一定存在严重的共线性。

表 11-10(b) 所示为模型中的变量、容忍度(容差)和方差膨胀因子(VIF)。该表显示, 变量“工作经验”与“年龄”的方差膨胀因子值分别为 2.882、2.880, 全部大于 2, 也说明存在共线性问题。

建议对可能存在共线性的变量作偏相关, 进一步分析在哪两个变量之间存在相关关系, 选代表性变量参与回归。

### 11.1.7 自动线性建模

#### 1. 自动线性建模概述

自动线性建模与一般线性回归分析一样也是研究自变量与因变量之间线性关系的一种二元或多元的统计分析方法。它对一般线性回归分析过程的功能进行了优化。在这个过程中既可以进行自动线性建模, 也可以像一般线性回归分析过程一样对线性回归中的一些算法参数进行手动设定建模, 但所需输入的参数要比【线性回归】过程中更少。相比而言, 它具有更加方便和快捷的特点。尤其在使用自动线性建模过程之前, 如果用户事先在“变量视图”的“角色”中对参与分析的数据文件中变量的“角色”属性进行预先的设定, 如将作为因变量的变量的角色定义为“目标”, 将自变量(预测变量)的“角色”定义为“输入”时, 则只要打开建模的数据文件, 调用【回归】菜单下的【自动线性建模】过程, 自动线性建模程序就会根据变量“角色”属性自动配置相应的变量为因变量和自变量, 用户无须做任何设置, 只要单击【自动线性建模】主菜单中的【确定】按钮, 自动线性建模程序就会根据系统默认设置自动进行建模, 在输出窗中得到自动线性建模的输出结果。自动线性建模过程还对一般线性回归分析过程的功能进行了拓展。在自动线性建模过程中, 提供了 4 种不同目的的建模方法, 除使用创建一个标准模型的方法可获取同使用【线性回归】过程一样的模型预测结果外, 它还提供了能提高预测准确性的 Boosting 法和预测稳定性的 Bagging 法, 以及适用于大型数据集的网络建模法。毫无疑问, 后 3 种方法, 使得线性建模的功能更加强大, 而这在【线性回归】过程中是不具备的。

#### 2. 自动线性建模中基于线性回归算法的 Boosting 法和 Bagging 法简介

在自动线性建模中用到的 Boosting 法和 Bagging 法是机器学习中经常用到的两种不同的集成算法。

##### (1) Boosting 法

在 SPSS 中, Boosting 法使用的是其家族中的基础算法 AdaBoost (Adaptive Boosting) 法 (Freund and Schapire 1996, Drucker 1997), 它是用来建立连续型因变量的增强模型的一种算法。它用整个原始数据集作为训练集来构建模型。系统默认建立的模型数量为 10, 该数量也可由用户自行设定。每次建立的模型称为基础模型。第一个是基础模型, 由于它与使用【线性回归】过程建立的模型的预测结果相一致, 因此也被称为标准模型。AdaBoost 算法的核心思想为: 初始化时对训练集中的每一个样品赋予相等的分析权重  $1/n$  ( $n$  为训练样本的样本量), 然后用加权线性回归的方法对训练集建立基础模型, 即得到预测函数。根据预测残差, 对具有较大绝对值预测残差的样品, 调高其分析权重; 同时, 对每个基础模型也赋予一定的模型权重, 模型预测效果较好的权重较大, 反之权重较小; 再用调整后的分析权重和模型权重, 重新对训练集建立基础模型, 重复这样的过程直到构建完指定模型数的最终模型为止。这样得到一个模型序列。由于后一个基础模型是建立在前一个模型的基础上的, 因此这些基础模型也被称为成分模型 (component models)。在 SPSS 中, AdaBoost 算法采用加权中位数的方法来获取组合模型 (ensemble model), 依此对新样品进行评分。

(2) Bagging 法

Bagging (Bootstrap Aggregating) 法又被称为自举聚合法。它建模时，每次使用从原始数据集中通过有放回的重重复抽样方式，获取与原始数据集相同容量的自举样本，作为训练集。系统默认的需要重复抽取的自举样本的数量为 10，也可由用户自行设定。Bagging 法中的频数权重是通过对二项分布中概率  $p$  的迭代运算得到的。这样，可在每个训练集上构建一个模型，再融合这些模型形成组合模型。对连续型因变量，组合规则是取这些模型预测值的均值来对新样品进行评分的。

3. 自动线性建模过程对变量的要求

在自动线性建模过程中，把因变量称为目标变量，而把自变量称为预测(输入)变量。自动线性建模过程对变量的要求是必须有一个因变量，并且至少有一个自变量。在默认情况下，预定义“角色”为“两者”或“无”的变量不能用作因变量或自变量。因变量，应是连续型(尺度)变量；对自变量的测度类型没有限制；分类(名义及有序)变量可用作线性模型中的因素变量，连续型变量可用作线性模型中的协变量。

需要注意的是：如果分类变量有 1000 个以上的类别，则自动线性建模程序不会运行，也不会创建任何模型。

4. 自动线性建模的过程

在数据编辑窗中打开一个数据文件，按【分析→回归→自动线性建模】顺序打开如图 11-13 所示的【自动线性建模】对话框【字段】选项卡。

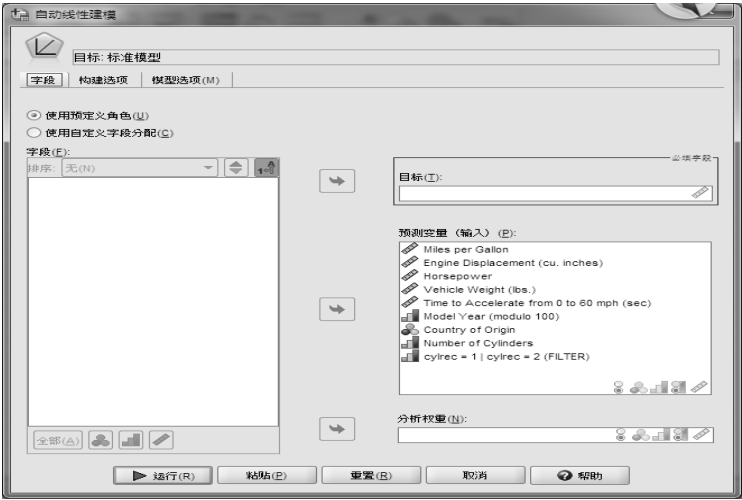


图 11-13 【自动线性建模】对话框【字段】选项卡

(1) 【字段】选项卡中的设定。

① 因变量、自变量的设定。如果在打开的数据文件中已定义因变量的角色为“目标”，自变量的角色为“输入”，则选择【使用预定义角色】，这是系统默认选项。如果在打开的数据文件中没有预先定义因变量的角色为“目标”，则选择【使用自定义字段分配】。此时，需要从左侧的【字段】名列表框中用手动选择因变量和自变量的方式，将因变量移入【目标】框中，将自变量移入【预测变量(输入)】框中。

② 分析权重变量的设定。如果对数据集中的每一个样品赋予的分析权重已建立了相应的数值型变量，则在左侧的【字段】名列表框中将其选中，并移到【分析权重】框中。

③ 对变量名进行排序。在左侧的【字段】名列表框中，单击【排序】按钮，有 3 种方式可对列表框中的变量名进行排序：

- **【无】**。系统默认选项，按导入和创建顺序排列变量名。
- **【字母数值】**。按英文字母的 ASCII 码和数值大小顺序排列变量名。
- **【测量(测度类型)】**。按名义、有序、尺度测度顺序排列变量名，而在同一个测度类型中，按导入和创建顺序排列字段名。

在【字段】名列表框下面有 4 个用来选择变量的按钮：

- **全部(A)**按钮。单击可选中【字段】名列表框中的所有变量。
- **人**按钮。单击可选中【字段】名列表框中所有名义变量。
- **上**按钮。单击可选中【字段】名列表框中所有有序变量。
- **尺**按钮。单击可选中【字段】名列表框中所有尺度变量。

(2) **【构建选项】**选项卡中的设定。在【自动线性建模】对话框【字段】选项卡中，单击【构建选项】按钮，打开如图 11-14 所示的【自动线性建模】对话框【构建选项】选项卡。

在【选择一项】框中有 5 个选项：

① **【目标】**。单击可得到如图 11-14 所示的【目标】选项卡。在这里需要确定【您的主要目标是什么？】，有以下 4 个目标可供选择：

- **【创建标准模型】**。是系统默认选项，构建一个使用预测变量预测因变量的标准模型。与下面 3 个选项建立的组合模型相比，标准模型更易解释且评分速度更快。
- **【增强模型精确性(Boosting)】**。使用 Boosting 法来构建组合模型。
- **【增强模型稳定性(Bagging)】**。使用 Bagging 法来构建组合模型。
- **【为大型数据集创建模型(需要 SPSS Statistics 服务器)】**。如果数据集非常大，无法构建上述任何模型，或希望建立增量模型时，选择本选项。选用本选项建模时，需要连接到 SPSS Statistics 服务器。它是一种将数据集拆分成单独的数据块来构建组合模型的方法。虽然建模用时要比标准模型短，但评分用时要比标准模型更长。

② **【基本】**。单击则弹出如图 11-15 所示的【基本】选项卡。



图 11-14 【构建选项】选项卡



图 11-15 【基本】选项卡

- **【自动准备数据】**选项。自动准备数据的目的是为了提高所建模型的预测性能。它是系

统默认选项。但当在【您的主要目标是什么?】选项中选择【为大型数据创建模型】时,本选项不可用。选择本选项,可以在内部自动对因变量和预测变量的值进行相应的数值转换,以便使得模型的预测能力最大化。程序将自动保存建模过程中用到的任何变量的数值转换规则,并将其应用于对新数据的评分。建立模型时,不再使用转换变量的原始数据。在默认情况下,程序将执行以下自动数据准备工作:

A **日期与时间处理**。每个日期型的预测变量将被转换成新的连续型预测变量,其中包含自参考日期(1970-01-01)起所经过的时间。每个时间型的预测变量被转换成新的连续型预测变量,其中包含自参考时间(00:00:00)起所经过的时间。

B **调整测量级别(调整测度类型)**。对不足 5 个不同值的连续型预测变量程序将会自动将其设成有序预测变量,而对于多于 10 个不同值的有序预测变量程序将会自动将其设成连续型预测变量。

C **离群值处理**。如果连续型预测变量的值位于临界值(平均值的 3 个标准差)之外,则将其作为离群值处理。

D **缺失值处理**。对名义预测变量的缺失值,将用训练分区的众数替换;对有序预测变量的缺失值,将用训练分区的中位数替换;对连续预测变量的缺失值,将用训练分区的平均值替换。

E **受监督的合并**。通过计算输入变量(自变量)与目标变量(因变量)间的关联关系来确定类似的类别,合并无显著差异(即  $p$  值大于 0.1)的类别。这样可以缩减自变量数,得到更简洁的模型。如果所有类别合并为一个类别,则模型中将不包含任何变量的原始和转换数据,因为它们没有值可用来作为预测变量。

● **【置信水平】**。用于输出结果为系数视图窗中计算模型系数的区间估计值的置信水平。该值需大于 0 且小于 100。系统默认值为 95。

③ **【模型选择】**。单击则弹出如图 11-16 所示的【模型选择】选项卡。

● **【模型选择方法】**。选择一种确定有哪些自变量可以进入线性模型的方法。可供选择的方法如下:

A **【包括所有预测变量】**。可将所有在字段选项卡上输入的预测变量全部选入线性模型。

B **【前向逐步选择(逐步向前选择)】**法。这是系统默认选项。选择该选项,则在开始时模型中没有任何自变量,然后,根据逐步选择自变量标准,在线性模型中添加满足条件要求的自变量,并剔除达到剔除标准的自变量,直到根据逐步选择标准不能再添加或剔除任一个自变量为止。

选择本选项需要再次选定其下的以下选项:

a. **【输入/删除标准(进入/剔除标准)】**选项。用来决定某个自变量是选入还是剔除出模型的统计量。它包括以下几个标准:

i. **【信息准则(AICC)】**。它是根据模型中给定训练集计算得到的似然估计值。用它可以避免过度复杂的模型。

ii. **【F 统计量】**。它是根据模型误差改进情况计算得到的一个检验统计量。如果选择该选项,



图 11-16 【模型选择】选项卡



需在【包括  $p$  值小于以下的效应】框中输入一个大于 0 小于 1 的数值，系统的默认值为“0.05”，这意味着在逐步建模的每步中将低于该指定阈值的所有  $p$  值中具有最小  $p$  值效应所对应的自变量将被添加到模型中。另外，还要在【删除  $p$  值大于此值的效应】框中输入一个大于 0 小于 1 的数值，该值须大于【包括  $p$  值小于以下的效应】框中所输入的值。系统的默认值为“0.10”，这意味着在逐步建模的每步中将模型中任何具有大于该指定阈值的  $p$  值的模型效应将被从模型中剔除。

iii. 【调整 R 方】。它是根据训练集计算得到的拟合度。用它可以避免过度复杂的模型。

iv. 【过度拟合防止准则(ASE)】。它是防止过度拟合集的拟合度(平均方差或 ASE)。防止过度拟合集是一个不用于训练模型且大约为原始数据集 30% 的随机子样本。

如果选择了【F 统计量】以外的其他标准，则在逐步建模的每步中将对应于选择标准具有最大正增长的效应添加到模型。对应于剔除标准具有减少情况的任何模型效应将被从模型中移除。

b. 【自定义最终模型中的最大效应数】。指定一个正整数作为最大效应数。逐步选择算法在达到指定的最大效应数时终止。如果逐步选择算法在具有指定最大效应数的某个步骤结束，则此算法将终止于当前效应集。在系统默认情况下，所有可用效应都将被输入到模型中。

c. 【自定义最大步骤数】。指定一个正整数作为最大步骤数。逐步选择算法在达到该指定步骤数后停止。在不输入任何值的情况下，此值默认为可用效应数的 3 倍。

d. 【最佳子集】。将检查“所有可能的”模型，或至少检查比向前逐步回归法中大一些的子集的可能模型，以选择满足相应标准的最佳子集。在【最佳子集选择】的【输入/删除标准】框中，有与【前向逐步选择】中相同的 3 个效应进出模型的标准，各标准的含义也同上。它将选择具有最大标准值的模型作为最佳模型。需要注意的是：与向前逐步选择法相比，最佳子集选择法涉及更多的计算。在与 Boosting 法、Bagging 法或大型数据集法配合执行最佳子集时，花费的时间要比使用向前逐步选择法构建标准模型时多得多。

④【整体】。单击则弹出如图 11-17 所示的【整体】选项卡。该选项卡中的设置决定了当【目标】选项卡选定 Boosting 法、Bagging 法或超大型数据集法时所发生的整体行为。它将忽略对选定目标不适用的选项。

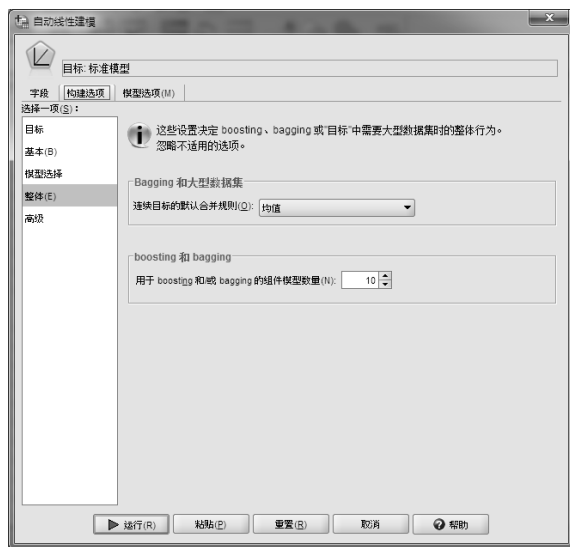


图 11-17 【整体】选项卡

• 【Bagging 和大型数据集】栏。在使用 Bagging 法和大型数据集法对整体评分时，此规则用于组合基础模型的预测值，以计算整体的得分值。

A. 【连续(型)目标的默认组合规则】。此选项中提供了两种方法来对连续型变量的整体预测值进行组合，一种是用基础模型的预测值的均值，另一种是用基础模型的预测值的中位数。

需要注意的是，如果在【目标选项卡】的【您的主要目标是什么？】选项中选定了【增强模型精确性】选项，则将忽略【连续目标的默认组合规则】的选择。对分类型因变量，Boosting 法始终使用加权众数来进行评分；而对连续型因变量，Boosting 法使用加权中位数来进行评分。

【Boosting 和 Bagging】栏。当在【目标】选项卡的【您的主要目标是什么？】选项中选定

了【增强模型精确性】或【增强模型稳定性】选项时，指定要构建的基础模型数。对应于 Bagging 法，该数为 Bootstrap 样本数。它应为正整数，系统默认值为 10。

⑤【高级】。单击则弹出如图 11-18 所示的【高级】选项卡。  
在该选项卡上只有一个【复制结果】选项。使用该选项可以用随机数设置随机种子，以便确定使用于过度拟合预防集中的记录。它可以指定一个正整数为随机数，也可以单击【生成】按钮来产生一个在 1~2147483647 之间(包括 1 和 2147483647)的伪随机整数。系统默认值为 54752075。

(3) 在【自动线性建模】对话框【字段】选项卡中，【模型选项】选项卡上的设定。单击【模型选项】按钮，打开如图 11-19 所示的【自动线性建模】对话框【模型选项】选项卡。



图 11-18 【高级】选项卡



图 11-19 【模型选项】选项卡

使用该选项卡可以将得分保存到活动数据集并将模型导出到外部文件。

①【保存预测值到数据集】。可在其后框中输入保存预测值的变量名。系统默认的变量名称是“预测值”。

②【导出模型】。将模型写入外部.zip 文件中。利用评分向导可以将得到的该模型文件的模型信息应用到其他数据文件的评分中。

在【导出模型】框中，可指定有效的唯一文件名。如果文件名为现有文件，则该文件将被覆盖。

单击【运行】按钮运行，则得到符合以上各设定选项的模型文件及预测值。

3. 评分向导

评分向导是实用程序中的一个过程，如图 11-20 所示。可以使用评分向导将由一个数据集创建的模型应用到另一个数据集，并生成得分。

(1) 按【实用程序→评分向导】顺序打开【评分向导】对话框，见图 11-21。

① 选择评分模型。单击【浏览】按钮导航到其他位置以选择评分模型文件。该模型文件可以是包含模型 PMML 的.xml 文件或.zip 存档文件。在【选择评分模型】列表框中，只显示扩展名为“.zip”或“.xml”的文件。列表中不显示文件扩展名。



图 11-20 【评分向导】菜单

图 11-20 【评分向导】菜单

② 模型详细信息。单击【选择评分模型】列表框中的文件名，则在【模型详细信息】框中显示所选模型的详细信息，包括建模方法、整体方法、应用、目标(因变量)、分割(拆分)和用于构建模型的预测变量。

由于必须读取模型文件才能获取该信息，因此在显示所选模型的此类信息之前可能会有延迟。如果.xml 文件或.zip 存档文件没有被识别为 SPSS Statistics 可以读取的模型，则将显示无法读取该文件的信息。

(2) 单击【下一步】按钮，则得到模型中所用到建模变量的信息，见图 11-22。



图 11-21 【评分向导】对话框



图 11-22 模型信息(一)

① 数据集字段。在其下拉列表中显示包含活动数据集中所有变量的名称。不能选择与相应模型变量的数据类型不匹配的变量。

② 模型字段。模型中使用的变量名称。

③ 角色。显示的角色可以是以下角色之一：

- 预测变量。该变量在模型中用作预测变量，即预测变量的值用于预测感兴趣因变量结果的值。
- 拆分。拆分变量的值用来定义分组，其中每个组单独评分。拆分变量值的每个唯一组合对应一个单独的分组。(注意：拆分变量只适用于一些模型。)
- 记录 ID。记录(个案)标识。
- 测量(测度类型)。模型中定义的变量的测度类型。对于测度类型会影响得分的模型，将使用模型中定义的测度类型，而不是活动数据集中定义的测度类型。
- 类型。模型中定义的数据类型。活动数据集中的数据类型必须与模型中的数据类型匹配。

数据类型可以是以下类型之一：

- A. 字符串。活动数据集中，数据类型为字符串的变量与模型中的字符串数据类型匹配。
- B. 数值。活动数据集中，显示格式不是日期或时间格式的数值变量要与模型中的数值数据类型匹配，其中包括 F(数值)、Dollar、Dot、Comma、E(科学计数法)和自定义货币格式。具有 Wkday(一周中的某天)和 Month(一年中的某月)格式的变量也被视为数值，而不是日期。对于一些模型类型，活动数据集中的日期和时间变量也被视为与模型中的数值数据类型匹配。

C. 日期。活动数据集中，显示格式包含日期但不包含时间的数值变量与模型中的日期类型匹配，其中包括 Date(dd-mm-yyyy)、Adate(mm/dd/yyyy)、Edate(dd.mm/yyyy)、Sdate(yyyy/mm/dd)和 Jdate(dddyyyy)。

D. 时间。活动数据集中，显示格式包含时间但不包含日期的数值变量与模型中的时间数据类型匹配，其中包括 Time (hh:mm:ss) 和 Dtime (dd hh:mm:ss)。

E. 时间戳。活动数据集中，显示格式同时包含日期和时间的数值变量与模型中的时间戳数据类型匹配。这对应于活动数据集中的 Datetime 格式 (dd-mm-yyyy hh:mm:ss)。

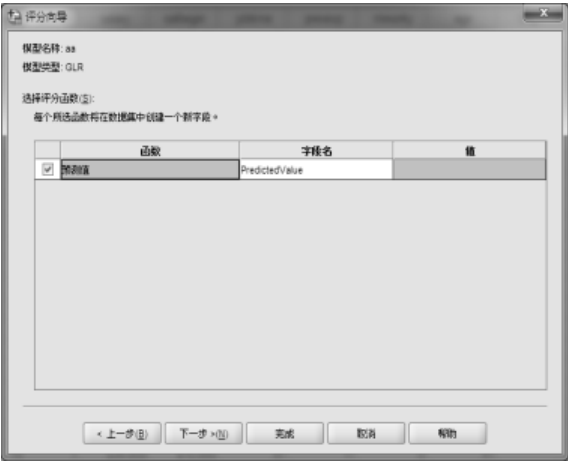


图 11-23 模型信息(二)

注意：除了字段名和类型以外，需要确保要评分的数据集中的实际数据值的记录方式与构建模型的数据集中的数据值的记录方式相同。例如，如果模型使用 Income 变量构建，后者将收入划分为 4 种类别，而活动数据集中的 IncomeCategory 则将收入划分为 6 种类别或 4 种不同的类别，这些变量实际上彼此并不匹配，结果得分将不可靠。

(3) 单击【模型信息(一)】中的【下一步】按钮，则得到模型中有关评分函数方面的信息，见图 11-23。

① 评分函数。有效的评分函数取决于模型。例如，因变量的预测值、预测值的概率或所选因变量值的概率等。每选一个评分函数，则在字段名中创建一个新变量。

下列值中的一个或多个在列表中是有效的：

- 预测值。感兴趣的因变量结果的预测值。除没有因变量的模型外，预测值在所有的其他模型中都是有效的。
- 预测值的概率。预测值的概率用正的比例值表示。它可用于具有分类因变量的大部分模型中。
- 所选值的概率。预测值的概率用正的比例值表示。从“值”列的下拉列表选择一个值。其可以使用的值由模型定义。它可用于具有分类因变量的大部分模型中。
- 置信度。与分类因变量的预测值有关的概率测度。对于二元 Logistic 回归、多项 Logistic 回归和 Naive Bayes 模型，其结果与预测值的概率相同。对于树和 Ruleset 模型，置信度可以被解释为预测类别的校正概率，而且始终比预测值的概率小。对于这些模型，置信度比预测值的概率更加可靠。
- 节点编号。树模型的预测终端的节点编号。
- 标准误。预测值的标准误。适用于尺度因变量的线性回归模型、一般线性模型和广义线性模型。它仅在模型文件中保存了协方差矩阵时有效。
- 累积风险。估计累积风险函数。在给定预测变量值的前提下，该值表示了在指定时间或该时间之前观察到事件的概率。
- 最近邻元素。最近邻元素的 ID。如果提供的话，该 ID 是案例标签变量的值，否则为个案编号。它仅适用于最近邻元素模型。
- 第 k 个最近邻元素。第 k 个最近邻元素的 ID。在“值”列中输入一个整数作为 k 值。如果提供的话，该 ID 是案例标签变量的值，否则则为个案编号。它仅适用于最近邻元素模型。
- 到最近邻元素的距离。到最近邻元素的距离。根据不同模型，将使用 Euclidean 或 block 距离。它仅适用于最近邻元素模型。
- 到第 k 个最近邻元素的距离。到第 k 个最近邻元素的距离。在“值”列中输入一个整数作

为  $k$  值。根据不同模型，将使用 Euclidean 或 block 距离。它仅适用于最近邻元素模型。

② 字段名称。每个选定的评分函数在活动数据集中保存一个新的变量，可以使用默认名称或输入新名称。如果活动数据集中已存在具有相同名称的变量，则它们将被替换。

③ 值。有关使用“值”设置函数的说明，请参见评分函数中的说明。

(4) 单击【模型信息(二)】中的【下一步】按钮，则得到如图 11-24 所示的【完成】选项卡。

这是向导中的最后一步，可以对活动数据集进行评分(或将所生成的命令语法粘贴到语法窗口。然后，可以在语法窗口中修改和/或保存所生成的命令语法)。

单击【完成】按钮，则在活动数据集中生成由所选评分函数所产生的新变量及其对应数值。



图 11-24 【完成】选项卡

#### 4. 自动线性建模的实例分析

**【例 3】** 使用数据文件 data11-01，以其中的 salbegin、prevexp、jobtime、educ、age 作为预测变量(自变量)，salary 作为因变量，用自动线性建模过程中的【逐步向前选择】法进行建模，并用所建模型在评分向导中对数据文件 data11-02 中的 salary 进行预测。

(1) 在数据编辑窗中，打开数据文件 data11-01。

(2) 按【分析→回归→自动线性建模】顺序打开如图 11-13 所示的【自动线性建模】对话框【字段】选项卡。

选择【使用自定义字段分配】选项，并将 salary 移入【目标】框中，将 salbegin、prevexp、jobtime、educ、age 移入【预测变量】框中。

(3) 单击【构建选项】选项卡，在【目标】中选择【创建标准模型】选项；在【基本】中选择【自动准备数据】选项；在【模型选择】中，建模的方法采用系统默认的【前向逐步(逐步向前)】法，在【前向逐步选择】的【输入/删除标准】中，选用【F 统计量】；其余采用系统默认值。

(4) 单击【模型选项】按钮，选择【将预测值保存到数据集】选项，字段名采用系统默认的“预测值”。选择【导出模型】选项，在文件名框中输入“F:\SPSS20.0 稿件\第 11 章回归分析\aa.zip”。单击【运行】按钮，则在当前数据文件中生成一系列变量名为“预测值”的新数据，它是用上述设定的方法，通过自动线性建模后所预测得到的当前工资的预测值，见图 11-25。建模中所有的信息都存放在 F:\SPSS20.0 稿件\第 11 章回归分析\aa.zip 文件中。

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	age	预测值
1	1	m	02/03/1952	15	3	\$57,000	\$27,000	98	144	0	52	\$56,065
2	2	m	05/23/1958	16	1	\$40,200	\$18,750	98	36	0	46	\$46,245
3	3	f	07/26/1929	12	1	\$21,450	\$12,000	98	381	0	74	\$22,650
4	4	f	04/15/1947	8	1	\$21,900	\$13,200	98	190	0	57	\$28,637
5	5	m	02/09/1955	15	1	\$45,000	\$21,000	98	138	0	49	\$44,744
6	6	m	08/22/1958	15	1	\$32,100	\$13,500	98	67	0	45	\$30,633
7	7	m	04/26/1956	15	1	\$36,000	\$18,750	98	114	0	48	\$40,480
8	8	f	05/06/1966	12	1	\$21,900	\$9,750	98	0	0	38	\$23,833

图 11-25 salary 的预测值

在输出窗中，得到如图 11-26 所示反映建模过程中用到的模型各种信息的缩略图。逐一双击这些缩略图可将它们在模型浏览器中激活，见图 11-27~图 11-32。

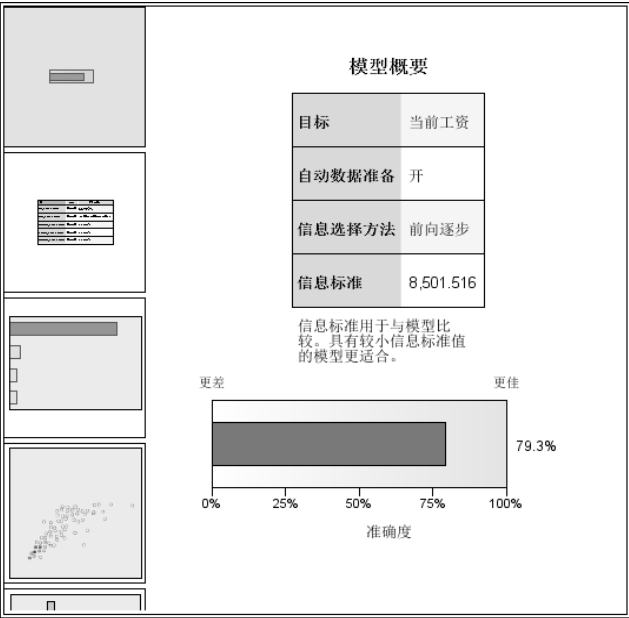


图 11-26 模型各种信息的缩略图

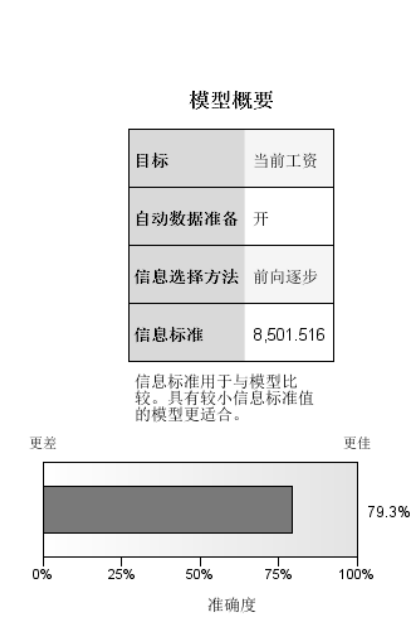


图 11-27 模型概要及其测度的准确度

图 11-27 所示为模型概要及其预测的准确度，由图可见，校正  $R^2$  值为 0.793，远大于 0.5，说明该模型可用，连同模型概要的信息标准值 8501.518 一起，还可对基于该数据集构建的上千个模型中的其他模型进行比较，以此选出最佳模型。从图中还可以看到，目标变量名为当前工资 (salary)，做了自动数据准备工作以及采用前向逐步法建模等信息。

图 11-28 所示为在模型构建之前自动对数据转换的准备情况。从图中可见，所有变量都进行过处理，对“受教育程度”变量进行过合并类别处理，其他变量都进行过去除离群值处理，并对“年龄”变量进行过替换缺失值的处理。

图 11-29 所示为预测变量在最终模型中的重要性顺序。排在图中最上方的变量在预测模型中对因变量的贡献最大。因此，预测变量的重要性依次为：“起始工资”、“受雇月数”、“受教育程度”和“年龄”。

自动数据准备		
目标：当前工资		
字段	角色	采取的操作
(age_transformed)	预测变量	去除离群值 替换 缺失值
(educ_transformed)	预测变量	合并类别使与目标的关联最大化
(jobtime_transformed)	预测变量	去除离群值
(prevexp_transformed)	预测变量	去除离群值
(salbegin_transformed)	预测变量	去除离群值

如果原始字段名为 X，则已转换字段显示为 (X transformed)。原始字段将从分析中排除，而已转换字段则会包括在分析中。

图 11-28 自动数据准备情况

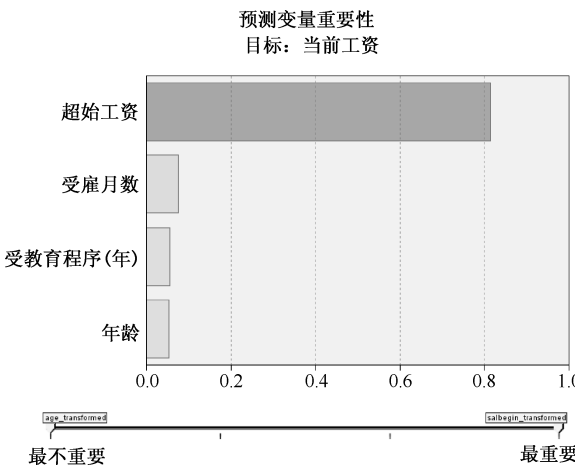


图 11-29 预测变量的重要性

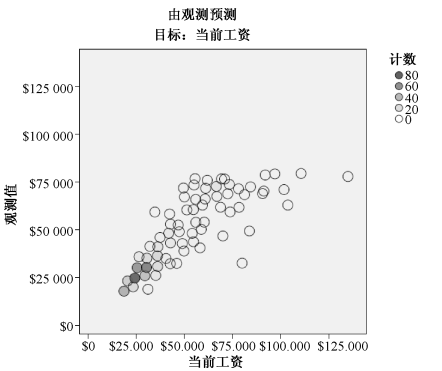
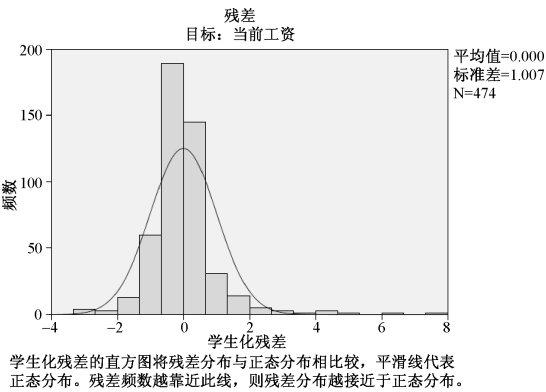


图 11-30 观测值与预测值散点图

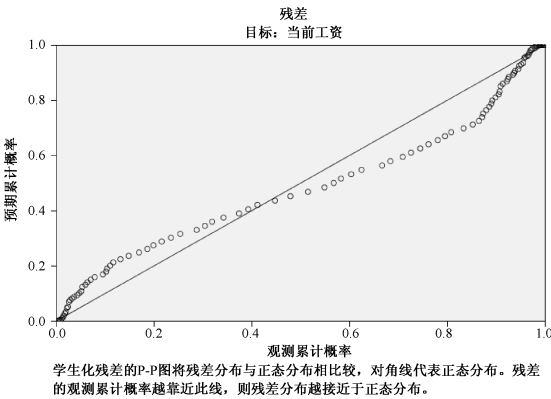
图 11-30 所示为观测值与预测值在直角坐标系中的散点分布情况。由于大部分点分布在 45° 直线周围，因此模型有较好的预测效果。

图 11-31 (a) 所示为残差分布的直方图，它与图中的正态曲线有较好的吻合度，表明残差服从或近似服从正态分布。

如果在显示本图的模型浏览器下方的【样式】下拉列表中选择【P-P 图】，则还可得到图 11-31 (b) 所示的残差的 P-P 图。



(a) 残差的直方图与正态曲线图



(b) 残差的 P-P 图

图 11-31 残差的直方图和 P-P 图

图 11-32 所示为离群值的分布情况。图中用 Cook 距离值(从上到下按由大到小的顺序做了排列)对因变量离群值大小、ID 号一一做了说明。

图 11-33 (a) 所示为各转换后的预测变量对目标变量的效应图。它按重要性的程度，由大到小、由上到下顺序排列。

如果在显示本图的模型浏览器下方的【样式】下拉列表中选择【表】，则还可得到图 11-33 (b) 所示的模型的方差分析结果及变量的重要性示意图。由于方差分析结果的显著性水平的值为 0.000，小于 0.05，因此它表明所得的线性模型有统计学上的显著性意义。

转换后的预测变量的重要性依次为：“起始工资”、“受雇月数”、“受教育程度”和“年龄”。

图 11-34 (a) 所示为转换后的各连续型变量及分类变量的各类在预测目标变量建模中的重要性大小及其在模型中的系数。各个预测变量的重要性的程度，也是按由大到小、由上到下顺序排列的。由于图中已清晰地标明，这里不再另行解释。在模型浏览器中用鼠标指向变量的线段，可以得到如图中所示的该变量在模型中的系数值、显著性值和重要性值。其他预测变量的值在这里不再一一列出，读者可自行验证。

如果在显示本图的模型浏览器下方的【样式】下拉列表中选择【表】，则还可得到图 11-34 (b) 所示的模型系数检验及变量的重要性表图。除“受教育程度”3 和 4 的系数没有显著性意义

离群值  
目标：当前工资

记录 ID	当前工资	Cook's 距离
29	\$135,000	0.281
343	\$103,500	0.106
18	\$103,750	0.098
446	\$100,000	0.093
32	\$110,625	0.079

图 11-32 离群值情况表

外，其余变量的系数均有显著性意义。图中一一给出了各变量重要性大小的测度值，它是图 11-34(a)的补充说明。

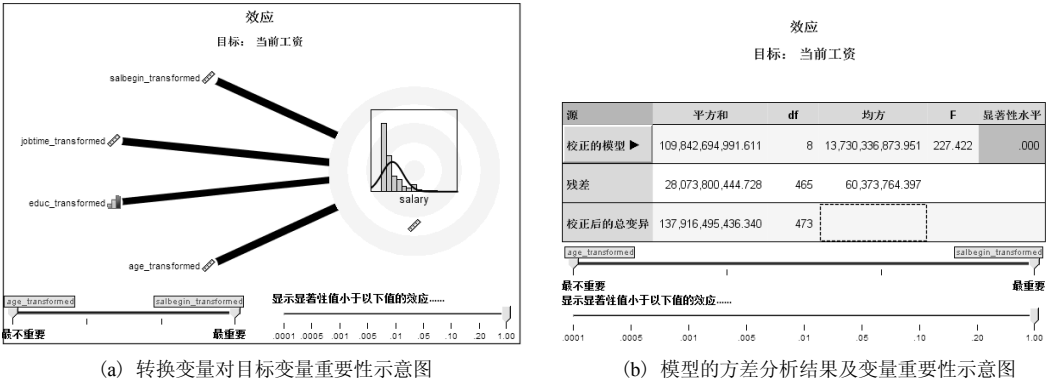


图 11-33 重要性示意图

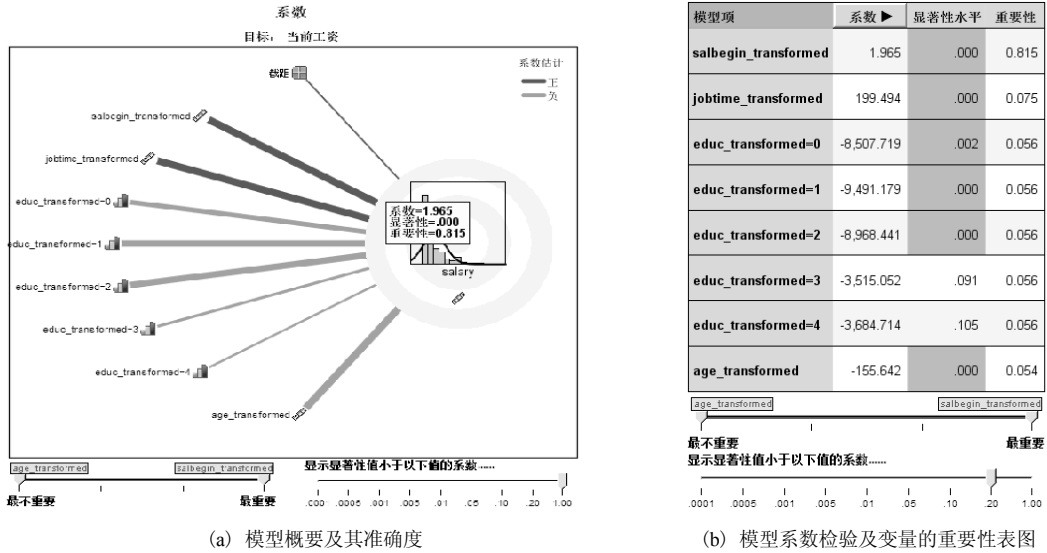


图 11-34

图 11-35 所示为根据有显著性效应的 4 个预测变量分别估计目标变量当前工资均值的线图。

图 11-36 所示为模型构建过程的汇总。模型用 F 统计量标准的逐步向前法构建。在每一步中有标记“✓”的变量意味着其进入模型。因此，在第一步中，起始工资变量进入模型，将鼠标光标指向第一步 F 检验的  $p$  值为 0.000 时，可得到此时对该变量检验的  $F$  值=1494.371,  $df1=1$ ,  $df2=472$ ，说明“起始工资”变量对模型具有显著性意义。在第二步中，受雇月数变量进入模型，对其检验的  $F$  值=28.686,  $df1=1$ ,  $df2=471$ 。在前五步中，所选 5 个自变量全部进入模型，在第六步中，由于过去经验变量的 F 检验  $p$  值为 0.291，达到剔除标准 0.10，因此，该变量从模型中剔除。

图 11-37 所示为使用自动线性模型过程来建模的全过程。它对整个自动线性建模过程中所做的一切工作进行了整体回顾。在目标中列出了自动线性建模的因变量为 salary，预测变量为 salbegin、prevexp、jobtime、educ、age，没有使用分区数据，目的是要创建标准模型，采用自动数据准备方法对原始数据进行建模前的预处理，设定的置信水平为 0.95，用逐步向前法建模，自变量进、出模型的标准采用 F 统计量的  $p$  值， $p$  值小于 0.05 的效应所对应的自变量进入模型，



$p$  值大于 0.10 的效应所对应的自变量从模型中剔除，最终模型中的预测变量为 salbegin、jobtime、educ、age 等信息。

(5) 打开数据文件 data11-02。

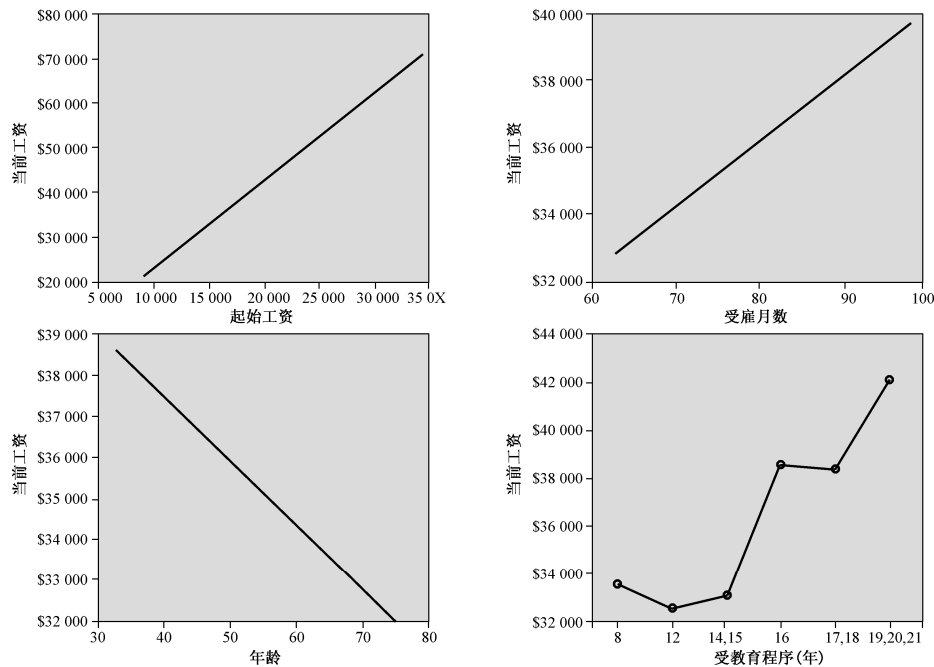


图 11-35 显示的前 10 个有显著效应的预测变量估计的目标变量的均值

	步骤					
	1	2	3	4	5	6
F 值的显著性	.000	.000	.000	.004	.040	.291
效应						
salbegin_transformed	✓	✓	✓	✓	✓	✓
jobtime_transformed		✓	✓	✓	✓	✓
prevexp_transformed			✓	✓	✓	
educ_transformed				✓	✓	✓
age_transformed					✓	✓

模型构建方法使用 F 统计量标准的前向逐步。  
选中标记意味着此步骤中效应在模型中。

图 11-36 模型构建汇总

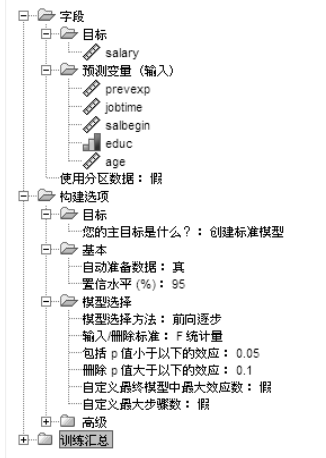


图 11-37 线性模型过程整体汇总

(6) 按【实用程序→评分向导】顺序打开【评分向导】对话框。单击【浏览】按钮，在【浏览评分模型】对话框中，选择评分模型文件

F:\SPSS20.0 稿件\第 11 章回归分析\aa.zip。  
单击【评分向导】对话框中的【完成】按钮，则在输出窗中得到反映模型中涉及的各项变量基本信息的一览表，见表 11-11。

从表 11-11 可见，模型中共有 5 个变

表 11-11 模型变量汇总表

SPSS Statistics 变量	模型变量					
	名称	标签	类型	宽度	角色	测量
salary	salary	预测值	数字	8	目标	连续
educ	educ	受教育程度(年)	数字	8	预测变量	序数
salbegin	salbegin	起始工资	数字	8	预测变量	连续
jobtime	jobtime	受雇月数	数字	8	预测变量	连续
age	age	年龄	数字	8	预测变量	连续

量: salary、educ、salbegin、jobtime、age, 均为数值型, 且宽度均为 8 位。其中, salary 是因变量, 其余 4 个变量为预测变量, 除 educ 变量为有序测度变量外, 其余变量均为连续型的等间隔测度变量。

用 aa.zip 中所建模型计算得到的因变量 salary 的预测值存放在当前数据文件新建的 PredictedValue 变量中, 见图 11-37。

id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	age	PredictedValue
200	m	02/13/1963	17	3	\$67,500	\$34,980	83	9	0	41	74559.28
201	m	05/08/1955	12	1	\$29,340	\$19,500	83	150	0	49	38282.50
202	m	03/17/1963	15	1	\$39,600	\$16,500	83	47	0	41	34156.83
203	m	03/17/1964	12	1	\$29,100	\$15,000	83	50	0	40	30842.95
204	m	10/21/1960	15	1	\$33,150	\$16,500	83	69	0	43	33845.54
205	m	06/22/1944	16	3	\$66,750	\$52,500	83	258	0	59	71927.36
206	m	05/22/1943	12	2	\$33,750	\$15,000	83	284	0	61	27574.46
207	m	02/15/1959	15	1	\$27,300	\$17,250	83	91	0	45	35007.64
208	f	11/28/1968	12	1	\$24,000	\$11,250	83	16	0	35	24254.23
209	f	01/14/1934	8	1	\$19,800	\$10,200	83	75	0	70	17727.46

图 11-37 salary 的 PredictedValue(预测值)

11.2 曲线估计

11.2.1 曲线回归概述

1. 一般概念

线性回归不能解决所有的问题。尽管有可能通过一些函数的转换, 在一定范围内将因、自变量之间的关系转换为线性关系, 但这种转换有可能导致更为复杂的计算或失真。

SPSS 提供了 11 种不同的曲线回归模型中。如果线性模型不能确定哪一种为最佳模型, 可以尝试选择曲线拟合的方法建立一个简单而又比较合适的模型。

2. 数据要求

(1) 自变量与因变量应为数值型变量。如果自变量以时间间隔测度, 则要求因变量也应是时间间隔测度的变量, 而且因、自变量使用的时间间隔和单位应是完全相同的。

(2) 模型的残差应该呈正态分布。如果选择了线性模型, 则因变量必须呈正态分布, 且所有的观测值应独立。

11.2.2 曲线回归过程

- (1) 按【分析→回归→曲线估计】顺序打开如图 11-38 所示的【曲线估计】对话框。
- (2) 在左侧的源变量框中选择一个或多个变量作为因变量, 送入【因变量】框中。
- (3) 在左侧的源变量框中选择自变量, 送入【自变量】的【变量】框中。如果因变量是时间间隔测度, 则直接选择【时间】选项。
- (4) 在【模型】栏中选择一个或多个拟合模型。各模型解释见表 11-12。



图 11-38 【曲线估计】对话框

表 11-12 不同模型的表示

模型名称	回归方程	相应的线性回归方程
线性	$y = b_0 + b_1t$	
二次项	$y = b_0 + b_1t + b_2t^2$	
复合	$y = b_0b_1^t$	$\ln y = \ln b_0 + (\ln b_1)t$
增长	$y = e^{(b_0 + b_1t)}$	$\ln y = b_0 + b_1t$
对数	$y = b_0 + b_1 \ln t$	
立方(三次)	$y = b_0 + b_1t + b_2t^2 + b_3t^3$	
S	$y = e^{b_0 + b_1/t}$	$\ln y = b_0 + b_1/t$
指数	$y = b_0 e^{b_1t}$	$\ln y = \ln b_0 + b_1t$
逆模型	$y = b_0 + b_1/t$	
幂	$y = b_0t^h$	$\ln y = \ln b_0 + b_1 \ln t$
Logistic	$y = 1/(1/u + b_0b_1^t)$	$\ln(1/y - 1/u) = \ln[b_0 + (\ln b_1)t]$

在表 11-12 中,  $t$  为时间或所指定的自变量;  $b_0$  为常数项;  $b_n$  为自变量第  $n$  次项的回归系数。如果选择了 Logistic 模型, 则模型中的  $u$  值必须是大于因变量最大值的正数。在【上限】框中输入这一上限值。

(5) 根据需要选择以下选项:

① 【在等式(方程)中包含常量】。

② 【显示 ANOVA 表格】。

③ 【根据模型绘图】。

④ 在源变量框中选择作为标识观测的变量, 送入【个案标签】框中。

(6) 单击【保存】按钮, 打开【曲线估计: 保存】对话框, 如图 11-39 所示。选择要保存在数据文件中的新变量, 包括预测值、残差、预测区间、显著性水平等变量。系统默认的新变量名与说明显示在输出窗口中。

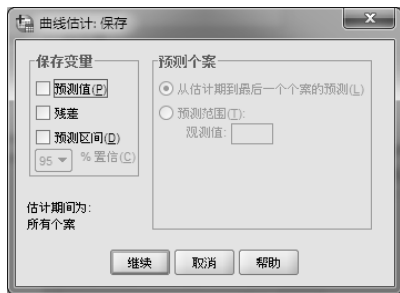


图 11-39 【曲线估计: 保存】对话框

① 【保存变量】栏。选项有因变量的【预测值】、【残差】值、【预测区间】; 在【置信】框中设置预测值的置信区间, 系统默认值为 95%。

② 【预测个案】栏。如果自变量为时间变量, 可以在该栏中指定一种超出当前数据时间序列范围的预测周期。

- 【从估计期到最后一个个案的预测内所有个案】都使用预先设定好的, 求出估计周期的预测值。估计周期和预测范围可以通过【数据】菜单中的【选择个案】命令设置。如果没有预先设置估计周期, 则计算时使用所有的观测。
- 【预测范围】。根据预先设定的周期, 对特定的数据、在指定的时间内进行预测。如果预测值的范围超出了时间序列的范围, 应选择该项, 并在【观测值】框中输入预测周期的末端值。

(7) 单击【确定】按钮提交运行, 在大多数情况下, 对变量之间关系的认识往往模糊不清, 需要先绘制散点图, 根据数据分布特点来确定应采用的模型。可以多指定几个模型进行拟合, 根据输出的统计量, 如  $R^2$  值, 再结合图形综合考虑, 确定最佳模型。

### 11.2.3 曲线回归分析实例

【例 4】用数据文件 data11-03 中的数据研究: 车重 weight 与每加仑千米数 mpg 之间的关系。

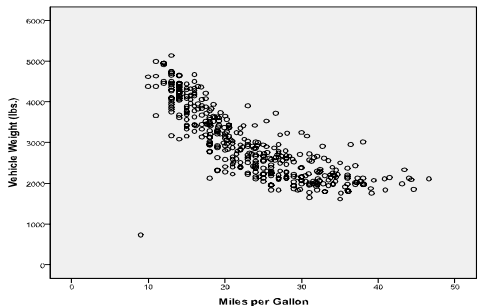


图 11-40 每加仑里程与车重散点图

1) 制作观测数据的散点图并初步选择模型

打开数据文件, 按【图形→旧对话框→散点/点状】顺序打开【散点图】对话框。以变量 mpg 作  $X$  轴, 变量 weight 作  $Y$  轴, 得到图 11-40 所示散点图。可见两个变量间呈明显的曲线关系。

2) 建立若干个曲线模型并进行比较

(1) 按【分析→回归→曲线估计】顺序打开主对话框。选择变量 mpg 作为因变量, weight 作为自变量。

(2) 在【模型】框中选择二次、三次与指数模型。

(3) 选择【显示 ANOVA 表格】、【根据模型绘图】及【在等式(方程)中包含常量】选项,

要求输出方差分析的结果和模型图形，方程包括常数项。单击【确定】按钮，提交系统执行。

3) 输出结果(见表 11-13~表 11-15 及图 11-41)

每组表对应 1 个模型的输出，每组表含 3 个子表：模型汇总、ANOVA(方差分析)、系数表。

表 11-13 二次项模型结果

模型汇总

R	R 方	调整 R 方	估计值的标准 误
.810	.656	.655	4.593

自变量为 Vehicle Weight (lbs.)。

ANOVA

	平方和	df	均方	F	Sig.
回归	15918.130	2	7959.065	377.209	.000
残差	8334.445	395	21.100		
总计	24252.575	397			

自变量为 Vehicle Weight (lbs.)。

系数

	未标准化系数		标准化系数	t	Sig.
	B	标准误	Beta		
Vehicle Weight (lbs.)	-.012	.002	-1.330	-6.094	.000
Vehicle Weight (lbs.) ** 2	7.597E-.007	.000	.528	2.419	.016
(常数)	52.540	3.030		17.337	.000

表 11-14 三次项模型结果

模型汇总

R	R 方	调整 R 方	估计值的标准误
.828	.686	.683	4.399

自变量为 Vehicle Weight (lbs.)。

ANOVA

	平方和	df	均方	F	Sig.
回归	16629.063	3	5543.021	286.476	.000
残差	7623.513	394	19.349		
总计	24252.575	397			

自变量为 Vehicle Weight (lbs.)。

系数

	未标准化系数		标准化系数	t	Sig.
	B	标准误	Beta		
Vehicle Weight (lbs.)	.033	.008	3.598	4.286	.000
Vehicle Weight (lbs.) ** 2	-1.434E-.005	.000	-9.968	-5.715	.000
Vehicle Weight (lbs.) ** 3	1.591E-.009	.000	5.655	.	.
(常数)	9.555	7.662		1.247	.213

4) 结果分析

模型汇总表列出复相关系数  $R$ 、判定系数  $R^2$ 、校正  $R^2$  值、标准误。

ANOVA 为方差分析结果：二次模型的  $F$  值为 377.209，三次模型的  $F$  值为 286.476；指数模型的  $F$  值为 957.936， $p$  值(表中的 Sig.)小于 0.0001。3 个回归方程均具有统计意义。

表 11-15 指数模型结果

模型汇总

R	R 方	调整 R 方	估计值的标准误
.841	.708	.707	.184

自变量为 Vehicle Weight (lbs.)。

ANOVA

	平方和	df	均方	F	Sig.
回归	32.405	1	32.405	957.936	.000
残差	13.396	396	.034		
总计	45.800	397			

自变量为 Vehicle Weight (lbs.)。

系数

	未标准化系数		标准化系数	t	Sig.
	B	标准误	Beta		
Vehicle Weight (lbs.)	.000	.000	-.841	-30.951	.000
(常数)	60.152	2.013		29.887	.000

因变量为 ln(Miles per Gallon)。

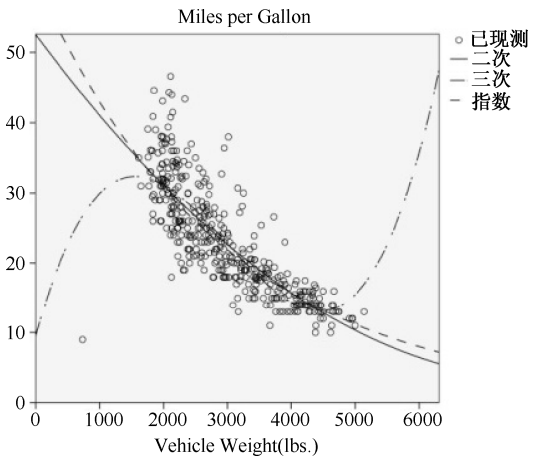


图 11-41 3 种模型的图形

系数表中显示回归系数  $B$ 、标准化回归系数  $Beta$  及其检验结果，由此得出各种模型的回归模型如下。

- 二次： $mpg = 52.540 - 0.012 \times weight + 7.597 \times 10^{-7} \times weight^2$ ；
- 三次： $mpg = 9.555 + 0.033 \times weight - 1.434 \times 10^{-5} \times weight^2 + 1.591 \times 10^{-9} \times weight^3$ ；
- 指数模型： $mpg = 60.152 \times 0.9996^{weight}$ 。

图 11-41 所示是 3 种模型的图形，似乎虚线即指数曲线对观测的拟合稍好一些。图形只对模型的取舍起辅助作用，最终的模型判定还是要通过对统计量的分析与研究进行。

① 比较 3 个模型的修正  $R^2$  值。指数模型的 Adjusted  $R^2 = 0.708$  最大；三次模型次之， $R^2 = 0.686$ ；二次模型的  $R^2 = 0.656$  最小。由此可以判断，拟合最好的是指数模型。

② 方差分析的  $F$  值概率均小于 0.001，因此比较  $F$  值。指数模型的  $F = 957.936$  最大；三次模型次之， $F = 377.209$ ；二次模型的  $F = 286.476$  最小。

通过以上判断得出最佳模型为  $\text{mpg} = 60.15 \times 0.9996^{\text{weight}}$ 。

注意：输出窗中表格中数据的小数显示位数设置为 5。

## 11.3 二项 Logistic 回归

在现实世界中，经常需要判断一些事情是否将要发生，如候选人是否会当选等。这类问题的特点是因变量只有两个值：发生(是)或者不发生(否)。这就要求建立的模型必须保证因变量的取值是 0 或 1。可是，大多数模型的因变量值常常处于一个实数集中，与因变量只有两个值的条件相悖。

本节介绍一种对因变量数据假设要求不高，并且可以用来预测具有两分特点的因变量概率的统计方法：二项(Binary)Logistic 回归模型。

当因变量具有两个以上的类别时，可以参考第 11.4 节介绍的用于分析多分变量的多项 Logistic 回归。

### 11.3.1 Logistic 回归模型

#### 1. Logistic 模型

在 Logistic 回归中可以直接预测观测相对于某一事件的发生概率，如果只有一个自变量，回归模型可以写作

$$\text{Prob}(\text{event}) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

式中， $b_1$  和  $b_0$  分别为自变量  $x$  的系数和常数； $e$  为自然常数。其曲线如图 11-42 所示。包含一个以上自变量的模型为

$$\text{Prob}(\text{event}) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

式中， $z = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$  ( $p$  为自变量的数量) 某一件事情不发生的概率为

$$\text{Prob}(\text{no event}) = 1 - \text{Prob}(\text{event})$$

可使用最大似然比法和迭代方法来建立 Logistic 模型。

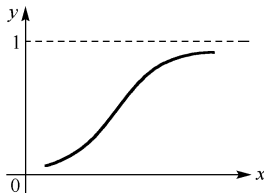


图 11-42 Logistic 回归曲线

#### 2. 数据要求

(1) 因变量应具二分特点，自变量可以是分类变量或等间隔测度的变量。如果自变量是分类变量，应为二分变量或被重新编码为指示变量。指示变量有两种编码方式。

① 指示变量编码方案。例如，当分类变量有 3 个水平(高、中、低)，就要创建两个新的指示变量。第一个变量：1 为低水平，0 为其他水平；第二个变量：1 为中间水平值，0 为其他水平值；高水平观测的两个变量值同时为 0。哪种水平为 0 值可任意决定。表 11-16 中参考类别的系数为 0。使用指示变量编码方法，只能比较每一类与参考类之间的效应差异。如果要比较每一类与整体的综合效果，应该选择如表 11-17 所示的编码方式。

② 背离编码方案。与编码方法一的区别仅仅在于新变量中最后一类被赋予-1 的编码值。利用这种编码方法, Logistic 回归系数展示每一类与各类综合效果的差异。见表 11-17, 对于每一个 SPSS 创建的新变量, 其系数代表了与综合效果之间的差异。注意, 最后一类的值应该是前两种系数之和并取负值。

表 11-16 指示变量编码方法

表 11-17 背离编码方法

Varvale (变量值)		Frequency (频数)	Paramenter coding (指示变量的编码设置)	
			(1)	(2)
Catacid (变量名称)	1.00	15	1	0
	2.00	20	0	1
	3.00	18	0	0

Varvale (变量值)		Frequency (频数)	Paramenter coding (指示变量的编码设置)	
			(1)	(2)
Catacid (变量名称)	1.00	15	1	0
	2.00	20	0	1
	3.00	18	-1	-1

(2) 自变量数据最好为多元正态分布, 自变量间的共线性会导致估计偏差。当观测分组完全依据分组变量时, 此方法十分有效; 当观测分组依据某连续型数值时(如根据智商得分可分高智商、低智商), 此方法会丢失连续型数据的信息, 应考虑线性模型。

3. Logistic 回归系数

为了理解 Logistic 回归系数的含义, 可以将回归方程改写为某一事件发生的几率。一个事件的几率被定义为它发生的可能性与不发生的可能性之比。例如, 抛一枚硬币后, 其正面向上的几率为 0.5/0.5=1; 从 52 张牌中抽出一张 A 的几率为 (4/52)/(48/52)=1/12, 这里不要将几率的含义与“概率”混淆, 其概率值为 4/52=1/13。

首先把 Logistic 方程写作几率的对数, 命名为 Logit:

$$\log \frac{\text{Prob(event)}}{\text{Prob(no event)}} = b_0 + b_1x_1 + \cdots + b_px_p$$

可以看出, Logistic 方程的回归系数可以解释为一个单位的自变量的变化所引起的几率的对数的改变值。由于理解几率要比理解几率的对数容易一些, 所以将 Logistic 方程式写为

$$\frac{\text{Prob(event)}}{\text{Prob(no event)}} = e^{b_0 + b_1x_1 + \cdots + b_px_p}$$

当第  $i$  个自变量发生一个单位的变化时, 几率的变化值为  $e^{(b_i)}$ 。自变量的系数为正值, 意味着事件发生的几率会增加,  $e^{(b_i)}$  的值大于 1; 如果自变量的系数为负值, 则意味着事件发生的几率会减少, 此值小于 1; 当  $b_i$  为 0 时, 此值等于 1。

4. 评价模型

建立模型后, 需要判断拟合的优劣。对大样本量的数据, 最好将数据分成两部分, 用一部分数据建立回归方程, 再将另一部分数据代入方程, 评定模型对数据的拟合情况。

(1) 系数检验。对于较大样本的系数检验, 使用基于卡方分布的 Wald 统计量。当自由度为 1 时, Wald 值为变量系数与其标准误比值的平方。对于两类以上的分类变量, Wald 统计量为  $W = B' V^{-1} B$ ,  $B$  为分类变量系数的极大似然估计向量,  $V^{-1}$  为变量系数渐近方差-协方差矩阵的逆矩阵。

Wald 统计量的弱点是当回归系数的绝对值变大时, 其标准误将发生更大的改变, 从而 Wald 值变得很小, 这将导致无法拒绝回归系数为 0 的零假设, 即认为变量的回归系数为 0。因此, 当变量的系数很大时, 不应依据 Wald 进行检验, 而应建立包含与不包含要检验的变量的两个模型, 利用对数似然比的变化值进行检验。可以选择 Backward LR 方式作为变量的选择方法。

## (2) 模型判别和模型校验。

① 模型判别。依据对事件发生的可能性的估计, 评估模型区分两组数据的能力。好的模型会将高概率的数值赋值给经常发生事件的观测, 不大可能发生的事件观测得到较小的概率值, 两种数据的概率不会发生重叠。

经常用来检查模型“判别”能力的指标为  $C$  统计量, 其值的范围为从 0.5~1。0.5 表示模型对观测的类别“判别”作用非常弱, 1 表示有强判别力。

SPSS 的 Logistic 回归过程, 先计算预测概率, 再利用 ROC 功能计算  $C$  统计量。

② 模型校验。评估观测概率、预测概率与整个概率之间的关系, 它对观测概率与预测概率之间的差异进行解释。当协变量配对的数量巨大, 且不能使用标准拟合度卡方检验时, 常用的检测方法 Hosmer 和 Lemeshow 卡方统计量非常有效。

计算 Hosmer 和 Lemeshow 卡方统计量, 先计算每一组中事件发生的实际观测数量与预测数量之间的差异, 然后按  $(\text{观测数量}-\text{预测数量})^2/\text{预测数量}$  计算, 卡方值为各分组中此值的和。

实际操作方法是根据估计观测数量的预测概率将观测分成数量大致相同的 10 个组, 观察观测到的数量与预测发生事件的数量以及预测不发生事件的数量之间的比较结果。卡方检测用来评价实际事件发生与预测事件发生之间的数量差别。使用这种鉴别方法时数据量要相当大, 以确保在大多数组别中至少有 5 个以上的观测, 同时所有的组别的预测值大于 1。

Hosmer 和 Lemeshow 卡方统计量的结果很大程度上与观测的分组情况有关。如果分组数很小, 得出的结果很可能与实际情况不符; 但如果观测数量很多, 则 Hosmer 和 Lemeshow 卡方统计量的结果也会变大。因此, 虽然 Hosmer 和 Lemeshow 卡方统计量在进行“模型校对”检测时是一种非常有效的方法, 但必须结合观测进行解释。

(3) 模型的拟合度是判别模型与样本的拟合优劣的统计量。利用已有的参数, 得出的观测结果的可能性称为“似然比”。似然比的值小于 1, 习惯上用对数似然比值乘以 -2 来度量模型对数据的拟合度, 记作  $-2\ln$ 。好的模型的似然比值较高, 其  $-2\ln$  值相对较小(如果模型 100%完美, 则似然比值等于 1,  $-2\ln$  值为 0)。似然比值的变化说明当变量进入与被剔除出模型时模型对数据拟合度方面的变化。

常用的 3 种卡方统计量分别为 Model、Block 和 Step。

① Model 统计量检验除常数项以外, 模型中所有变量系数为零的假设。卡方值为当前模型的与模型中只包含常数项的  $-2\ln\text{-likelihood}$  之差。

② Block 卡方值为当前模型与后一组变量进入模型后的  $-2\ln\text{-likelihood}$  值之差。如果选择了多组变量, 那么 Block 卡方值用来对最后一组变量系数为 0 的零假设进行检验。

③ Step 卡方值是在建立模型的过程中, 当前与下一步  $-2\ln$  之间的差值。它用来对最后一个加入模型的变量系数为 0 的零假设进行检验。

## (4) 评价包含所有变量模型的拟合效果。

① Cox & Snell  $R^2$  & Nagelkerke  $R^2$  统计量。与线性模型中的  $R^2$  相似, 是对 Logistic 模型变异中可解释部分的量化

$$\text{Cox \& Snell } R^2 = 1 - \left( \frac{L(0)}{L(B)} \right)^{\frac{2}{N}}$$

式中,  $L(0)$  为方程中只包含常数项时的似然比值;  $L(B)$  为方程包含设定变量时的似然比值;  $N$  为样本量; Cox & Snell  $R^2$  统计量最大值不可能为 1。1991 年, Nagelkerke 修改了 Cox & Snell  $R^2$

统计量,使其最大值可以为 1,即

$$\text{Nagelkerke } R^2 = \frac{R^2}{R^2_{\max}}$$

式中,  $R^2_{\max} = 1 - [L(0)]^{\frac{2}{N}}$ 。反映了由回归方程解释的变异百分比。

② 偏差。对于每个观测,其偏差值为  $(-2\lg \text{ 预测概率})^{0.5}$ 。例如,某男性患者预测其没有患恶性淋巴结的概率为 0.80 时,其偏差为  $-\sqrt{-2\lg 0.8} = -0.668$ 。大样本数据的偏差往往近似正态分布。偏差较大暗示模型拟合数据欠佳。学生化残差与偏差之差可以用来检测非常态数据。以  $P_i$  为第  $i$  个观测的预测概率,残差的 Logit 值计算公式为

$$\frac{\text{residual}_i}{P_i(1 - P_i)}$$

(5) 影响点的查找。

① 杠杆值 (Leverage) 检测哪些观测对预测值产生影响较大。与线性回归不同,在 Logistic 回归中杠杆值依据因变量得分和设计矩阵。其值在 0~1 之间,它们的均值为  $P/N$ ,其中  $P$  为模型中估计参数的个数(包括常数项), $N$  为观测的个数。对于那些预测概率值大于 0.9 或小于 0.1 观测来说,虽然观测具有影响力,但其杠杆值较小。

② Cook 距离用来检测观测的影响力。说明如果删除了一个观测后对这个观测残差的影响和对其他观测残差的影响。

$$\text{Cook 距离} \quad D_i = \frac{Z_i^2 \times h_i}{1 - h_i}$$

式中,  $Z_i$  为标准化残差;  $h_i$  为杠杆值。

③ DfBeta 统计量,即当删除一个观测后 Logistic 系数的变化值。

$$\text{DfBeta}(b_1^{(i)}) = b_1 - b_1^{(i)}$$

式中,  $b_1$  为当所有观测包括在模型中时的系数值;  $b_1^{(i)}$  为排除第  $i$  个观测后的系数值。

较大的变化值暗示应对此观测给予重新检查。

(6) 与线性回归相同,交互项可以作为新变量参与回归分析并包含在回归方程中。

11.3.2 二项 Logistic 回归过程

(1) 按【分析→回归→二元(二项)Logistic】顺序打开如图 11-43 所示的对话框。



图 11-43 【Logistic 回归】对话框

(2) 选择一个具有两分属性的变量作为因变量送入【因变量】框。

(3) 选择一个或多个变量为协变量送入【协变量】框。也可以同时选择两个和多个变量作为交互项,单击【>a\*b>】按钮,将它们送入【协变量】框。

(4) 在【方法】框中确定一种自变量进入模型的方式。

① 【进入】。自变量全部进入模型。

② 【向前: 条件】。向前逐步选择法。将变量剔除出模型的依据是,条件参数估计的似然比统计量的概率值。

③ 【向前: LR】。依据最大偏似然估计所得的似然比统计量的概率值,向前逐步选择变量。



- ④ **【向前：Wald】**。依据 Wald 统计量的概率值向前逐步选择变量。
- ⑤ **【向后：条件】**。根据条件参数估计似然比统计量的概率值，向后逐步剔除变量。
- ⑥ **【向后：LR】**。依据最大偏似然估计值统计量的概率值向后逐步剔除变量。
- ⑦ **【向后：Wald】**。根据 Wald 统计量的概率向后逐步剔除变量。

(5) **【选择变量】**框。根据指定变量的取值范围，确定参与分析的观测。在源变量框中选择一个变量，送入**【选择变量】**框中。单击**【规则】**按钮，打开**【Logistic 回归：设置规则】**对话框，如图 11-44 所示，设置选择观测值的标准。例如，要选择  $\text{time} = 100\text{s}$  的变量，那么选择 time 变量后在算数操作符框中选择**【等于】**，然后在**【值】**框中输入“100”。

SPSS 会将选择的观测值与非选择的观测值的计算结果全部显示出来。

(6) 单击**【分类】**按钮，打开如图 11-45 所示的**【Logistic 回归：定义分类变量】**对话框，设置处理分类变量的方式。



图 11-44 **【Logistic 回归：设置规则】**对话框



图 11-45 **【Logistic 回归：定义分类变量】**对话框

- ① **【协变量】**框中包含了在主对话框中已经选择好的全部协变量及交互项。
- ② **【分类协变量】**框中列出了所选择的分类变量，其后面的括号中显示的是各组间的对比方案，字符串变量将自动进入**【分类协变量】**框。
- ③ **【更改对比】**栏。设置分类协变量中各类水平的对比方式。
- **【指示符】**。指示出是否同属于参考分类，参考分类在对比矩阵中以一横排“0”表示。
  - **【简单】**。每种分类的预测变量(参考类别除外)效应都与参考类别效应进行比较。
  - **【差值】**选项。除第一类外，每类的预测变量效应都与其前所有各分类的平均效应进行比较，也称逆 Helmert 对比。
  - **【Helmert】**。除最后一类外，每类的预测变量效应都与其后所有各类的平均效应进行比较。
  - **【重复】**。除第一类外，每类的预测变量效应都与其前一种分类的效应进行比较。
  - **【多项式】**。对角多项式对比，要求每类水平相同，仅适用于数字型变量。
  - **【偏差】**。每类的预测变量(参考分类除外)效应与总体效应进行比较。
  - **【参考类别】**。如果选择了**【偏差】**、**【简单】**、**【指示符】**对比方式，可选择**【第一个】**或**【最后一个】**选项，指定分类变量的第一类或最后一类作为参考类。

如果改变了**【更改对比】**的设置，则单击**【更改】**按钮以示对选项的确定。

(7) 在主对话框中单击**【保存】**按钮，打开如图 11-46 所示**【Logistic 回归：保存】**对话框，选择在数据窗中保存的新变量。

- ① 在**【预测值】**栏中选择**【概率】**选项，则新变量存取的是每个观测发生特定事件的预

测概率;选择【组成员】选项,则新变量存取的是依据预测概率得到的每个观测的预测分组值。

②【影响】栏。存取每一个观测对预测值影响的统计量的值,包括【Cook 距离】、【杠杆值】和【DfBeta】统计量。

③【残差】栏。选取需要保存的残差类型。其可选项有【未(非)标准化】、【Logit】、【学生化】、【标准化】和【偏差】。

④【将模型信息输出到 XML 文件】栏。指定输出模型信息到 XML 格式的文件,单击【浏览】按钮,确定保存位置和文件名。选择【包含协方差矩阵】选项则输出中还包括协方差矩阵。

(8) 单击【选项】按钮,打开如图 11-47 所示的【Logistic 回归: 选项】对话框,设置各种检测参数。

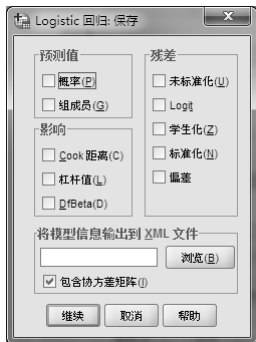


图 11-46 【Logistic 回归: 保存】对话框



图 11-47 【Logistic 回归: 选项】对话框

①【统计量和图】栏。选择要求输出的统计量与图表。

- 【分类图】。因变量的预测值与观测值的分类直方图。
- 【Hosmer-Lemeshow 拟合(优)度】。统计量。
- 【个案残差列表】。对每个观测输出非标准化残差、预测概率、观测的实际与预测分组水平。
  - A. 【外离群值】□【标准差】。在空格处输入一个正数,表示要求只输出那些标准化残差值大于输入值的观测值的各种统计量。系统默认值为 2。
  - B. 【所有个案】选项。要求输出所有观测的各种统计量。
  - C. 【估计值的相关性(系数)】选项。要求输出方程中各变量估计参数的相关系数矩阵。
  - D. 【迭代历史记录】选项。要求进行参数估计时,每一步迭代都输出相关系数和对数似然比值。
  - E. 【CI for exp(B)】选项。输出 exp(B) 的置信区间。选择此项需在其后框中处输入 1~99 的数值。系统默认值为 95。

②【输出】栏设置输出范围。【在每个步骤中】选项要求对每步计算过程输出表、统计量和图形;【在最后一个步骤中】选项,只要求输出最终方程的表格、统计量和图形。

③【步进概率】栏。用来设置变量进入模型及从模型中剔除的判定标准。如果变量的概率值小于【进入】后框中的设置值,那么此变量进入模型中,如果其概率值大于【删除】后框中的设置值,变量会被从方程式中剔除。【进入】的默认值为 0.05,【删除】的默认值为 0.10。此处的设置值必须为正数,而且【进入】值必须小于【删除】值。

④【分类标准值】框。用来设置系统划分观测类别的辨别值。大于设置值的观测被归于一组中,反之观测将被归于另一组中。其值的范围为 0.01~0.99,系统默认值为 0.5。

⑤【最大迭代次数】框。定义输出最大的迭代步数。

⑥【在模型中包括常数值】。用来设定模型中包括常数项。

11.3.3 二项 Logistic 回归分析实例

【例 5】 数据文件 data11-04 中是乳腺癌患者的数据。利用 age 年龄、pathscat 扩散等级、pathsize 肿瘤尺寸变量,建立一个预测因变量 ln\_yesno 癌变部位的淋巴结是否含有癌细胞的模型。

- 1) 操作步骤
- (1) 按【分析→回归→二元 Logistic】顺序打开其对话框。
- (2) 将变量 ln\_yesno 选入【因变量】框,将变量 pathsize、age、pathscat 作为自变量依次选入【协变量】框。
- (3) 打开【Logistic 回归: 定义分类变量】对话框,将变量 pathscat 选入【分类协变量】框中,在【更改对比】框中选择【指示符】方式。对扩散等级变量 pathscat 重新编码为指示变量。
- (4) 打开【Logistic 回归: 选项】对话框,选择【分类图】、【Hosmer-Lemeshow 拟合(优度)】和【CI for exp(B)】选项,在【输出】栏中选择选项【在最后一个步骤中】。
- (5) 在【Logistic 回归: 保存】对话框中选择【概率】、【组成员】选项,以便观察哪些患者属于淋巴结有癌细胞可能性较大;选择【杠杆值】选项,通过杠杆值查找影响点;选择【标准化】选项,观察标准化残差以便使用图形对模型进行诊断。

其他选项为 SPSS 的默认选项,单击【确定】按钮提交运行。

- 2) 输出结果(见表 11-18~表 11-28)
- 表 11-18 所示为在计算过程中的观测数量和缺失值的数量,以及它们所占的百分比。
- 表 11-19(a)所示为因变量变量的编码,表 11-19(b)所示是自变量中的分类变量在模型中根据指示变量编码方案所生成的新变量表。新生成的变量名称为参数编码(1) (pathscat(1))与参数编码(2) (pathscat(2))。

表 11-18 观测简表

案例处理汇总		
未加权的案例 <sup>a</sup>	N	百分比
选定案例 包括在分析中	1121	92.9
缺失案例	86	7.1
总计	1207	100.0
未选定的案例	0	.0
总计	1207	100.0

a. 如果权重有效,请参见分类表以获得案例总数。

表 11-19 因变量与分类变量代码表

因变量编码		分类变量编码				
		频率	参数编码			
初始值	内部值		(1)	(2)		
No	0	Pathological Tumor Size	<= 2 cm	826	1.000	.000
		(Categories)	2-5 cm	283	.000	1.000
			> 5 cm	12	.000	.000
Yes	1					

(a)

(b)

- 表 11-20 所示是在模型中没有自变量的情况下其初始的预测结果,预测正确的个案数为 860,预测错误的个案数为 261,预测正确率为  $860/(860+261)=76.7\%$ 。
- 表 11-21 所示为在初始状态下,模型中只有常量一项  $B=-1.192$ ,其标准误为 0.071, Wals 统计量值为 284.699,自由度  $df=1$ ,  $Sig.=0.000$ ,预测错误率的几率为 0.303。

表 11-20 没有自变量进入模型的初始状态

分类表 <sup>a,b</sup>		已预测		
		Lymph Nodes?		百分比校正
已观测	Lymph Nodes?	No	Yes	
		860	0	100.0
步骤 0	No			
	Yes	261	0	.0
总计百分比				76.7

a. 模型中包括常量。  
b. 切割值为 .500

表 11-21 初始模型的 Wals 检验

方程中的变量						
	B	S.E.	Wals	df	Sig.	Exp (B)
步骤 0 常量	-1.192	.071	284.699	1	.000	.303

表 11-22 所示为拟合起步前模型外的变量的卡方检验。所有单个变量 Sig. 值均小于 0.01, 4 个变量的总卡方检验 Sig. 值也小于 0.01, 故所有变量均有资格进入模型。

表 11-23 所示为 3 种常用的卡方统计量。因为拟合方法选择的是默认的【全部进入】法, 只有一步完成包含常数项与 5 个变量的模型的拟合, 所以模型的第一步、拟合过程块和模型的卡方值全部相同。如果采用的是逐步回归, 增加变量, 一步计算后的 Sig. 值小于 0.05, 那么说明增加变量后的方程有意义; 剔除一个变量的一步计算后, 如果 Sig. 值大于 0.10, 那么说明剔除变量后的方程仍然有意义。

表 11-22 起始模型外的变量

不在方程中的变量		得分	df	Sig.
步骤 0 变量	pathsize	49.161	1	.000
	age	31.793	1	.000
	pathscat	34.262	2	.000
	pathscat(1)	26.897	1	.000
	pathscat(2)	19.449	1	.000
	总统计量	67.558	4	.000

表 11-23 第一步模型系数卡方检验表

模型系数的综合检验			
	卡方	df	Sig.
步骤 1	64.897	4	.000
步骤块	64.897	4	.000
模型	64.897	4	.000

表 11-24 所示为模型拟合优度统计量。表中的-2ll 值为 1151.770。此值较大, 说明模型对数据的拟合度不理想。接下来是 Cox & Snell  $R^2$  和 Nagelkerke  $R^2$  统计量, 其值分别为 0.056、0.085, 值太小, 说明能由方程解释的回归变异太少, 拟合效果不佳。

表 11-25 中的 Hosmer-Lemeshow 是拟合统计量, 其零假设为方程对数据的拟合良好。本例 Sig. > 0.05, 无法拒绝零假设。这与表 11-23 的结论有差异, 故需要参考其他统计量。

表 11-24 最终模型的拟合优度检验

模型汇总			
步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	1151.770 <sup>a</sup>	.056	.085

a. 因为参数估计的更改范围小于 .001, 所以估计在迭代次数 4 处终止。

表 11-25 Hosmer-Lemeshow 检验表

= Hosmer 和 Lemeshow 检验 =			
步骤	卡方	df	Sig.
1	8.545	8	.382

表 11-26 所示为以概率值为模型对淋巴结中是否含有肿瘤细胞进行 Hosmer-Lemeshow 检验的列联表。依据对观测的预测(淋巴结中是否含有肿瘤细胞)概率, 它们被分为大致相等的 10 个组, “总计” 栏是每组观测总数。由于将具有相同值的观测组合在一起, 所以每组的观测数并非精确地相等。第 2、3 栏分别为观测到的和预测的淋巴结中不包含肿瘤细胞的数量, 第 4、5 栏分别为观测到的和预测的淋巴结中包含肿瘤细胞的数量。例如, 在第一组中的 114 个观测中实际有 14 个观测(预测为 12.3)到淋巴结中包含肿瘤细胞, 100 个观测(预测接近 101.66)到淋巴结中不包含肿瘤细胞, 其余各行的预测值与观测值都比较接近。

表 11-27 是以 0.5 作为淋巴结阳性与阴性(淋巴结中包含、不包含肿瘤细胞)分界线得出的预测值与实际数据的比较表。从表中可看到, 846 名淋巴结中没有肿瘤细胞的观测对象被正确地预测, 正确率为 98.4%; 同时 246 名包含肿瘤细胞的患者被错误地预测为淋巴结中不包含恶性肿瘤细胞, 正确率为仅为 5.7%; 总的正确判断率为 76.8%。显然, 这个回归方程不能在实际中应用。据此可以估计淋巴结中发现癌细胞的概率为

$$\text{prob(淋巴结中有癌细胞)} = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

表 11-28 所示为模型中的各变量的相关统计量。根据表中各变量的系数 “B”, 可以写出

$$z = -0.398 + 0.424\text{pathsize} - 0.025\text{age} - 0.185\text{pathscat}(1) - 0.307\text{pathscat}(2)$$

在当前数据文件中，生成预测概率和分类等新变量，见图 11-48。图中的新变量“PRE\_1”是预测概率，“PGR\_1”是预测分类。可以看到 PRE\_1 小于 0.5 的分到没有癌细胞的淋巴转移组，预测概率大于 0.5 的预测为有淋巴转移。

表 11-26 Hosmer-Lemeshow 检验的列联表

Hosmer 和 Lemeshow 检验的随机性表						
		Lymph Nodes? = No		Lymph Nodes? = Yes		总计
		已观测	期望值	已观测	期望值	
步骤 1	1	100	101.658	14	12.342	114
	2	102	95.735	9	15.265	111
	3	96	94.001	16	17.999	112
	4	88	90.962	23	20.038	111
	5	92	90.386	21	22.614	113
	6	86	87.109	26	24.891	112
	7	84	83.486	27	27.514	111
	8	74	81.573	39	31.427	113
	9	72	75.218	40	36.782	112
	10	66	59.871	46	52.129	112

表 11-27 最终观测分类表

分类表 <sup>a</sup>						
		已预测				
		Lymph Nodes?		百分比校正		
		No	Yes			
步骤 1	已预测	Lymph Nodes?	No	846	14	98.4
			Yes	246	15	5.7
		总计百分比				76.8

a. 切割值为 .500

表 11-28 最终模型统计量

方程中的变量									
		B	S.E.	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
a	pathsize	.424	.131	10.487	1	.001	1.528	1.182	1.975
	age	-.025	.006	18.282	1	.000	.976	.965	.987
	pathscat			.548	2	.760			
	pathscat(1)	-.185	.846	.048	1	.827	.831	.158	4.362
	pathscat(2)	-.307	.728	.178	1	.673	.736	.176	3.066
	常量	-.398	1.042	.146	1	.702	.671		

在步骤 1 中输入的变量: pathsize, age, pathscat.

3) 作散点图

按【图形→旧对话框→散点图/点图】顺序单击菜单项，在弹出的【散点图/点图】选项卡中，选择简单分布图标，单击【定义】按钮，则弹出【简单散点图】对话框，将数据文件中的新变量杠杆值“[LEV\_1]”定义为 Y 轴，将“ID”定义为 X 轴，单击【确定】按钮运行，则在输出窗中得到杠杆值的散点图，见图 11-49。

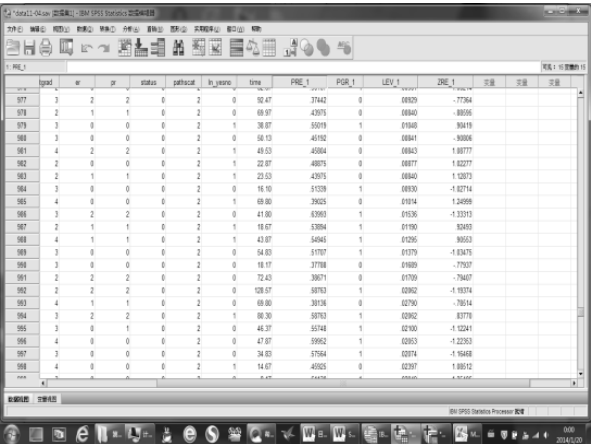


图 11-48 新变量：预测的概率与分类

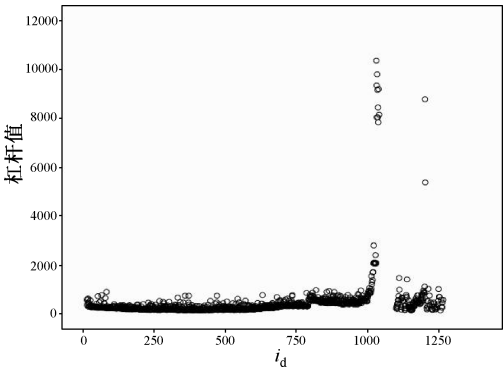


图 11-49 查找影响点的杠杆值散点图

由图 11-49 中可见, 杠杆值较大的对模型影响较大。双击该图进入图形编辑状态, 对认为是影响点的离群点双击, 在右键菜单中选择【显示数据标签】标出影响点的 ID 号。可以据此对这几个观测进行深入研究。

**【例 6】** 计算某年龄为 60 岁, pathsize 肿瘤大小为 1cm, 扩散等级 pathscat 为 2 的患者扩散到淋巴的概率。

注意: 根据表 11-19 的编码方式, 本例 pathscat(1) 的值为 0, pathscat(2) 的值为 1。

计算:  $z = -0.398 + 0.424 \times 1 - 0.025 \times 60 - 0.185 \times 0 - 0.307 \times 1 = -1.781$

其淋巴结中发现癌细胞的概率  $p = e^{-1.781} / (1 + e^{-1.781}) \approx 0.144 = 14.4\%$

在大多数情况下, 如果此值小于 0.5, 基本可以预测事件不会发生, 大于 0.5 则反之。结合查看图 11-48 也可以大致推测此人淋巴结中含有癌细胞的可能性不大。但由于模型可靠程度太低, 结论只能作为参考。

## 11.4 多分变量 Logistic 回归

因变量为多水平分类变量的情况在医学领域中常见, 如在某一药物试验中, 动物服药后的状态是 A(变量值为 1)、B(值为 2)、C(值为 3)或是 D(值为 4)等。当因变量为多(水平)分类变量时, 可以使用多分变量 Logistic 回归的方法建立回归模型。

### 11.4.1 多分变量 Logistic 回归的概念

#### 1. Logistic 回归基本概念

对于因变量的  $k-1$  个水平, 每个水平一个回归方程, 每个水平的因变量概率值为  $0 \sim 1$ 。自变量是连续变量或计数变量(非标称变量)的, 可以用 Logistic 回归方法对因变量的概率值建立回归模型。回归曲线为典型的 S 形, 见图 11-42。例如, 为了使得电影市场更加贴近观众, 可以使用电影观众的年龄、性别以及他们更喜欢观看的电影类型来建立多分变量 Logistic 回归, 预测常看电影的观众更喜爱哪种类型的影片。

Logistic 模型写为

$$\lg \frac{p(\text{event})}{1 - p(\text{event})} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

式中,  $b_0$  为常数项;  $b_1 \sim b_p$  为 Logistic 模型的回归系数, 是 Logistic 回归的估计参数;  $x_1 \sim x_p$  为自变量。模型的左侧称为 Logit, 是事件发生几率的自然对数值。

如果因变量具有  $j$  类可能性, 第  $i$  类的模型为

$$\lg \frac{p(\text{category}_i)}{1 - p(\text{category}_j)} = b_{i0} + b_{i1} x_1 + b_{i2} x_2 + \cdots + b_{ip} x_p$$

这样, 对于每一个 Logit 模型都将获得一组系数。例如, 如果因变量具有 3 种分类, 将会获得两组非零参数。

Logistic 回归方程的另一种形式为

$$p = e^y / (1 + e^y)$$

式中,  $y = a + \sum b_i x_i$  或  $y = \ln[p / (1 - p)]$ 。

通过变换可以得出  $p$  与变量  $x_i$  之间的数学表达式

$$p = \frac{e^{(a + \sum b_i x_i)}}{1 + e^{(a + \sum b_i x_i)}}$$

## 2. 数据要求

因变量应该是分类变量，自变量为因素变量与协变量（因素变量必须为分类变量，协变量必须是连续型变量）。

## 3. 模型检验

(1) 拟合检验。

① Pearson 卡方统计量在多维表中检测观测频数与预测频数间的差异。其公式为

$$\chi^2 = \sum_{\text{所有单元格}} \frac{(\text{观测数量} - \text{预测数量})^2}{\text{预测数量}}$$

其值越大，显著性概率越低，模型拟合效果越不好。

② Deviance 卡方是另一个检测模型拟合度的指标。其公式为

$$\chi^2 = 2 \sum_{\text{所有单元格}} \text{观测数量} \times \ln \frac{\text{观测数量}}{\text{预测数量}}$$

如果模型对数据拟合得好，对数似然比的差值就小，显著性水平值越大。大样本数据的 Deviance 卡方与 Pearson 卡方的值相近。

(2) 伪  $R^2$  统计量。在 Logistic 回归模型中使用 Cox & Snell、Nagelkerke 和 Mc Fadden 统计量。前两个已经介绍过，这里介绍 McFadden 统计量。其公式为

$$R_{\text{Mc Fadden}}^2 = \frac{l(0) - l(B)}{l(0)}$$

式中， $l(B)$  为模型中对数似然比的核； $l(0)$  为仅包含截距的模型的对数似然比的核。

(3) 观测—控制量的“配对”研究。它是一种利用现有观测数据研究那些很难发生的事件或是数据难以收集的事件。

例如，汽车销售公司为了分析购买奔驰汽车客户的特点，一般不得不收集大量的客户信息来确保分析的有效性，而利用观测—控制量的“配对”研究就可以不必收集很多购买了奔驰汽车的客户信息。这里，观测为已有的购买了奔驰汽车的客户信息，控制量是那些没有购买奔驰汽车的客户信息。观测和控制量通过它们之间共有的年龄和性别进行配对。

① 对于包含  $k$  对观测和控制量的数据，“经历”某种事件的 Logit 模型可以写成

$$\lg(P_i) = a_k + \sum_{i=1}^p b_i x_i$$

式中， $a_k$  为根据配对变量值得到的第  $k$  对变量的“风险”； $x_1 \sim x_p$  为未配对自变量的值； $b_i$  为第  $i$  个配对自变量的 Logistic 回归系数； $P_i$  是事件的概率。

② 创建“差异变量”。SPSS 分析过程可以对满足特殊要求的一对一的变量数据进行分析。在配对分析中，观测样本的样本量必须和与其配对的控制样本的样本量相同，并且差异变量必须是配对的观测与控制量间的差异。如果配对数多于 1 个，则差异是平均值间的差。

现有 56 对母亲的数据，其中一半的数据具有婴儿出生时较低体重的特点，另一半没有这样的特点，它们之间根据年龄（配对变量）配对。

其中的变量包括 lwt（怀孕前的体重）、age（年龄）、race（种族，1：白种人，2：黑色人种，

3: 其他人种); smoke(怀孕期间是否吸烟, 1: 吸烟, 0: 不吸烟); ptd(以前是否分娩, 0: 没有, 1: 有过); 以及 ui(子宫过敏, 1: 是, 0: 否)。

表 11-29 包含了配对后各变量之间的“差异”。虽然, 看起来计算观测和控制量之间的差异比较容易, 但是当使用的分类变量超过两类时就会产生一些困难。在 SPSS 的 Logistic 回归过程中, 类似的分类变量必须事先定义为因素变量。在进行观测-控制量配对分析时, 必须创建新变量替代分类变量, 并找到那些新变量之间的差异。

考虑一个简单的例子, 种族变量具有 3 个值, 所以需要使用两个变量表示它们。如果使用编码 1 表示参考类, 必须创建两个新变量 race1 和 race2, 其编码如表 11-30 所示。计算变量 race1 和 race2 之间的差值作为其他类。

表 11-29 配对变量之间的差异

	low	lwt	age	race	smoke	ptd	ui	race1	race2
观测	1	101	14	3	1	1	0	0	1
控制量	0	135	14	1	0	0	0	0	0
差 异	1	- 34		X	1	1	0	0	1

表 11-30 种族的编码方式

	race1	race2
White	0	0
Black	1	0
Other	0	1

③ 数据文件格式。观测-控制量的“配对”研究数据文件变量安排为: 因变量、配对变量、观测 1、控制量 1、差异变量 1、观测 2、控制量 2、差异变量 2、……、观测 *n*、控制量 *n*、差异变量 *n*。本例数据文件中应建立如下变量 low(因变量、其值应全部为 1-0=1)、age(配对变量)、caslwt(观测 lwt)、conlwt(控制量 lwt)、diflwt(lwt 的差异变量)、cassmoke(观测 smoke)、consSmoke(控制量 smoke)、difsmoke(smoke 的差异变量)……

最终将变量 low 设置为因变量, 差异变量设置为协变量。

注意: 如果交互项存在, 首先必须创建交互项, 然后计算它们之间的差异。在数据文件中的每一个观测应该包含因变量、配对变量、控制变量、差异变量, 以便将它们应用到相关的交互项分析中。对所有的观测来说, 因变量必须设置为一个常量, 并且所有的差异变量必须设置为协变量(Covariates)。配对变量不能够作为主效应进入模型中, 这是由于它们之间的差异为零。

11.4.2 多分变量 Logistic 回归过程

(1) 按【分析→回归→多项 Logistic】顺序打开如图 11-50 所示的主对话框。

(2) 在左侧的源变量框中选择一个多分类变量作为因变量送入【因变量】框中。一般情况下, 多项 Logistic 过程默认因变量的最后一类作为参考类, 如果要重新进行设置, 可单击【参考类别】按钮进行设置, 如图 11-51 所示。

① 【参考类别】栏。设置参考类。可选择【第一类别】或【最后类别】项分别将第一类或最后一类作为参考类; 【设定】选项后面由用户设置除第一和最后类别以外的其他类别作为参考类。

② 【类别顺序】栏。选择【升序】项, 将分类变量中值最小的类设为第一类, 值最大的类设为最后一类; 选择【降序】项, 则与【升序】顺序完全相反。

(3) 在源变量框中选择一个或多个分组变量送入【因子】框中。

(4) 在源变量框中选择一个或多个连续型变量作为协变量送入【协变量】框中。

(5) 单击【保存】按钮打开【多项 Logistic 回归: 保存】对话框, 见图 11-52。

① 在【保存变量】栏中, 选择要生成并保存到当前数据文件中的新变量:

- 【估计响应概率】。估计观测进入因变量各组的响应概率值。
- 【预测类别】。预测的观测分类。





图 11-50 【多项 Logistic 回归】主对话框



图 11-51 【多项 Logistic 回归：参考类别】对话框



图 11-52 【多项 Logistic 回归：保存】对话框

- 【预测类别概率】。预测观测各分类结果的概率。
- 【实际类别概率】。实际分类的概率值。

② 在【将模型信息输出到 XML 文件】栏选择文件，可将模型信息保存到外部 XML 格式文件中。

- 存储路径和文件名可以通过单击【浏览】按钮打开的对话框中指定，也可以直接输入。
- 选择【包含协方差矩阵】选项，可要求在输出的外部文件中包括协方差矩阵。

(6) 单击【条件(标准)】按钮打开【多项 Logistic 回归：收敛性准则】对话框，见图 11-53，设置模型拟合过程结束的判定标准。

① 在【迭代】栏中设置迭代停止的判定标准。

- 在【最大迭代】框中，设置最大迭代数。它必须为小于等于 100 的正整数。系统默认值为 100。
- 在【最大步骤对分(最大逐步二分法)】框中，输入使用逐步二分法的最大步数。系统默认值为 5。
- 在【对数似然性收敛性】框中，可以给定一个正数来设置对数似然比收敛值。当回归过程中的对数似然比大于此值时，迭代过程将停止。系统默认值为 0。
- 在【参数收敛性】框中，可设置收敛参数。在模型拟合过程中，如果绝对变化值或相对变化值大于等于此值时，迭代过程将停止。系统默认值为 0.000001。
- 在【为每一项打印迭代历史记录】项中，设置输出迭代过程的步距。系统默认值为 1。
- 在【从迭代中向前检查数据点的分离情况】项中，设置检查迭代过程开始值。系统默认值为 20。

② 在【Delta】框中输入小于 1 的非负值，此值会出现现在交叉表的空单元中。系统默认值为 0。这将有助于稳定算法、阻止估计偏差。在【奇异性容许误差】下拉列表中选择检验单一性的容忍度值。系统默认值为 0.00000001。

(7) 在主对话框中，单击【模型】按钮打开【多项 Logistic 回归：模型】对话框，见图 11-54。在【因子与协变量】框中包含协变量和因素变量。



图 11-53 【多项 Logistic 回归：收敛性准则】对话框



图 11-54 【多项 Logistic 回归：模型】对话框

① 在【指定模型】栏中指定模型。

- 【主效应】选项。选定主效应选项，则在模型中只包括协变量和因素变量的主效应。
- 【全因子】选项。选定全因子选项，则在模型中包含所有的主效应以及它们之间可能的交互效应。

• 【设定/步进式】(自定义)选项。选定该项，用户可自行设定模型中包括的主效应和交互效应。

② 以下选项只有在选定【设定/步进式】选项后生效。

- 在【建立项】栏的下拉列表中可选择一种效应类型，包括【交互】、【主效应】、【所有二阶】、【所有三阶】、【所有四阶】、【所有五阶】。
- 【强制输入项】框。选择强制出现在方程中的效应项进入此框。
- 【步进项】(逐步进入项)框。选择要逐步加入或剔除出模型的效应项进入此框。
- 在【步进法】(逐步进入方法)下拉列表中可以选各效应项逐步进入方程的方法，包括【向前进入】法、【向后去(剔)除】法、【向前步进(逐步向前选择)】法、【向后步进(逐步向后选择)】法。

③ 【在模型中包含截距】选项，选择它则要求在模型中包含截距项。

(8) 主对话框中单击【统计量】按钮，打开如图 11-55 所示的对话框。选择在输出窗显示的统计量。

① 【个案处理摘要】选项。给出分类变量综合信息。

② 在【模型】栏中选择的模型统计量包括：

• 【伪方】。在输出窗中将显示 Cox & Snell、Nagelkerke  $R^2$  和 McFadden  $R^2$  等统计量。

• 【步骤摘要】(逐步筛选摘要)。在输出窗中将显示每一步变量进入或被剔除出方程时的效应表。只有在模型对话框中指定用逐步法建模的情况下，才会生成此表。

• 【模型拟合度信息】。在输出窗中将显示模型拟合优度信息。

• 【信息标准】。在输出窗中将显示有关模型的判定标准信息。

• 【单元格可能性】(单元格概率)。在输出窗中将显示观测与期望频数表(带有残差)、协变量比率和响应分类。

• 【分类表】。输出每一类中观测和预测的分类表。

• 【拟合度】。输出 Pearson 卡方和似然比卡方统计量。

• 【单调性测量(度)】。输出表中包括和谐对数、不和谐的对数和节点数，和谐指数 C，以及 Somers'D、Goodman、Kruskal's Gamma、Kendall's tau-a 等统计量。

③ 【参数】栏。指定要输出与模型参数有关的统计量。

• 【估计】。输出模型的各种参数估计值，包括由用户设置的置信区间。

• 【似然比检验】。自动输出整个模型的检验统计量和模型的偏效应的似然比检验统计量。

• 【渐进(近)相关】。输出参数估计的相关阵。

• 【渐进(近)协方差】。输出参数估计的协方差矩阵。

• 【置信区间(%)】框。设置置信区间。系统默认值为 95。

④ 【定义子总体】栏。在此可选择因素变量和协变量的子集，以便定义协变量模式，用于单元概率和拟合优度检验。

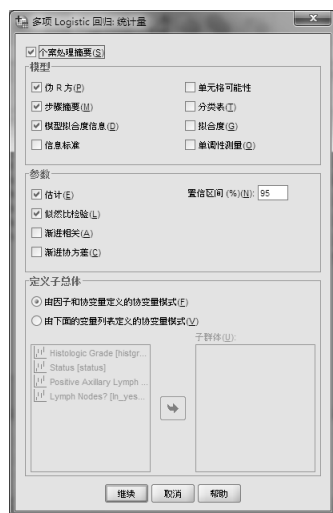


图 11-55 【多项 Logistic 回归：统计量】对话框

- 【由因子和协变量定义的协变量模式】。对所有因子变量和协变量进行拟合优度卡方检验。此为默认选项。
- 【由下面的变量列表定义的协变量模式】。在左下角的框中选择希望计算拟合优度卡方检验统计量的变量，将其送入右下角的【子群体】(总体)框中。

11.4.3 多分变量 Logistic 回归分析实例

【例 7】 数据文件 data11-05 中是 1992 年美国总统选举的数据，用变量 sex 预测选民投票结果 pres92。

1) 操作步骤

- (1) 打开数据文件 data11-05。按【分析→回归→多项 Logistic】顺序打开相应对话框。
- (2) 将投票变量 pres92 作为因变量选入【因变量】框中；将变量 sex 性别作为因素变量选入【因子】框中；在【多项 Logistic 回归：统计量】对话框中选择【参数】栏下的【估计】复选项。
- (3) 其他选项为 SPSS 的默认选项，单击【确定】按钮提交运算。

2) 输出结果(见表 11-31~表 11-35)

Pres92 的值使用值标签。

表 11-31 基本统计量小结

案例处理摘要		N	边际百分比
VOTE FOR CLINTON, BUSH, PEROT	Bush	661	35.8%
	Perot	278	15.1%
	Clinton	908	49.2%
RESPONDENTS SEX	male	804	43.5%
	female	1043	56.5%
有效		1847	100.0%
缺失		0	
总计		1847	
子总体		2	

表 11-32 模型拟合信息

模型	模型拟合标准	似然比检验		
	-2 倍对数似然值	卡方	df	显著水平
仅截距	61.209			
最终	27.343	33.866	2	.000

表 11-31 所示为基本统计量，包括：投给布什、克林顿和帕洛特的票数、百分比，投票人性别比例。从表中可见，投票人中，女性比例为 56.5%，高于男性的 43.5%；投给候选人克林顿的票数最多，占投票者中的百分比为 49.2%，其次为投给候选人布什的票数，占投票者中的百分比为 35.8%，投给帕洛特的票最少。

表 11-33 伪 R 方表

伪 R 方	
Cox 和 Snell	.018
Nagelkerke	.021
McFadden	.009

表 11-32 所示是模型拟合信息、最终方程的有效性检验，Sig.值为 0.000，小于 0.01，因此方程有效。

表 11-33 所示是伪 R 方值表，由于所有 R 方值均小于 0.022，远小于 1，所以回归效果不佳。

表 11-34 所示为似然比统计量检测每一个变量对方程的影响，sex 变量的 Sig.值小于 0.01，说明变量 sex 对方程具有显著性意义。

表 11-35 中 Wald 统计量的 Sig.值全部小于 0.001，因此可以将 Logit 模型写为

$$G1 = \lg \frac{P(\text{布什})}{P(\text{克林顿})} = -0.5 + 0.433(\text{sex}) \quad (\text{sex})$$

$$G2 = \lg \frac{P(\text{帕洛特})}{P(\text{克林顿})} = -1.51 + 0.715(\text{sex}) \quad (\text{sex})$$

表 11-34 似然比卡方检验

似然比检验				
	模型拟合标准	似然比检验		
	简化后的模型的-2 倍对数似然值	卡方	df	显著水平
截距	27.343 <sup>a</sup>	.000	0	
sex	61.209	33.866	2	.000

卡方统计量是最终模型与简化后模型之间在-2 倍对数似然值中的差值。通过从最终模型中省略效应而形成简化后的模型。零假设就是该效应的所有参数均为 0。

a. 因为省略效应不会增加自由度，所以此简化后的模型等同于最终模型。

表 11-35 模型参数估计

参数估计									
VOTE FOR CLINTON, BUSH, PEROT <sup>a</sup>		B	标准误差	Wald	df	显著水平	Exp(B)	Exp(B) 的置信区间 95%	
Bush	截距	-.501	.068	54.067	1	.000			
	[sex=1]	.433	.104	17.422	1	.000	1.543	1.258	1.891
	[sex=2]	0 <sup>b</sup>	.	.	0	.	.	.	.
Perot	截距	-1.511	.098	235.703	1	.000			
	[sex=1]	.715	.139	26.572	1	.000	2.044	1.558	2.682
	[sex=2]	0 <sup>b</sup>	.	.	0	.	.	.	.

a. 参考类别是: Clinton。

b. 因为此参数冗余，所以将其设为零。

由于男性 sex 值为 1，女性值为 0。因此简化了女性的 Logit 模型。例如，第一个截距-0.5 解释为女性选布什的概率与选择克林顿概率之比的自然对数；第二个截距-1.51 解释为女性选帕洛特的概率与选择克林顿概率之比的自然对数。变量 sex 的系数说明了 Logit 和性别之间的关系。因为所有的系数为正值并有显著意义。可以看出，男性选布什和帕洛特的可能性要比女性大得多。

表 11-35 中的系数描述了使用克林顿作为参照类别时不同性别的两个 Logit 模型，同时获得了候选人之间的对比结果。也可以将布什与帕洛特进行对比，根据  $\lg(a/b) = \lg a - \lg b$ ，可以推出

$$\lg \frac{p(\text{布什})}{p(\text{帕洛特})} = \lg \frac{p(\text{布什})}{p(\text{克林顿})} - \lg \frac{p(\text{帕洛特})}{p(\text{克林顿})}$$

查看参数 exp(B) 可知，男性选民选择布什的几率是女性选民的 1.54 倍(布什与克林顿进行比较)，选择帕洛特的概率是女选民的 2.04 倍(帕洛特与克林顿作比较)。

性别变量为什么能对投谁的票有很好的判断能力呢？为分析不同性别对投票对象起决定作用是否是因为不同性别选民的受教育的年限不同，可以增加 educ 受教育程度作为协变量到模型中将变量 educ 送入【协变量】栏，其他操作同上，运行结果见表 11-36。

从表 11-36 可以看出，增加了 educ，性别变量的系数改变很小。Wald 检验的零假设是回归系数均为 0。受教育年限的 Wald 统计量的 Sig. 值全都大于 0.05，说明无法拒绝该变量系数为 0 的假设。因此也可以尝试将与受教育程度相关的学位变量作为因素来拟合模型。因为 educ 是连续变量，而学位变量是分类变量，可以作为因素变量。

表 11-36 变量 educ 作为协变量的模型参数及检验结果

参数估计									
VOTE FOR CLINTON, BUSH, PEROT <sup>a</sup>		B	标准误差	Wald	df	显著水平	Exp(B)	Exp(B) 的置信区间 95%	
Bush	截距	-.702	.259	7.318	1	.007			
	educ	.015	.018	.656	1	.418	1.015	.979	1.051
	[sex=1]	.428	.104	16.970	1	.000	1.535	1.252	1.881
	[sex=2]	0 <sup>b</sup>	.	.	0	.	.	.	.
Perot	截距	-1.894	.353	28.859	1	.000			
	educ	.027	.024	1.248	1	.264	1.028	.980	1.078
	[sex=1]	.715	.139	26.396	1	.000	2.043	1.556	2.684
	[sex=2]	0 <sup>b</sup>	.	.	0	.	.	.	.

a. 参考类别是: Clinton。

b. 因为此参数冗余，所以将其设为零。

3) 将性别变量和学位变量都作为因素变量作分析(结果见表 11-37~表 11-39)

表 11-37 中的卡方值是排除因素变量与最终模型的两个-2lg Likelihood 的差值的卡方。检验结果 Sig. 值小于 0.001，说明最终模型成立。

表 11-38 所示是因素变量性别 sex、学历 degree 在最终模型中的似然比卡方检验结果。这是根据某个效应剔除出模型后的 $-2\ln$  值的变化情况进行的检验。其零假设为某变量被从模型中剔除后该统计量没有变化。从表中的 Sig. 值得出: sex 和 degree 剔除出模型后,  $-2\ln$  变化显著, 拒绝性别和学历在模型中系数为 0 的假设。

表 11-37 模型拟合信息

模型	模型拟合标准	似然比检验		
	-2 倍对数似然值	卡方	df	显著水平
仅截距	178.457			
最终	103.601	74.856	10	.000

表 11-38 似然比卡方检验

效应	模型拟合标准	似然比检验		
	简化后的模型的-2 倍对数似然值	卡方	df	显著水平
截距	103.601 <sup>a</sup>	.000	0	.
sex	140.753	37.153	2	.000
degree	144.590	40.990	8	.000

卡方统计量是最终模型与简化后模型之间在 -2 倍对数似然值中的差值。通过从最终模型中省略效应而形成简化后的模型。零假设就是该效应的所有参数均为 0。

a. 因为省略效应不会增加自由度, 所以此简化后的模型等同于最终模型。

表 11-39 加入学位变量后参数估计及其检验

参数估计									
VOTE FOR CLINTON, BUSH, PEROT <sup>a</sup>		B	标准误	Wald	df	显著水平	Exp(B)	Exp(B) 的置信区间 95%	
Bush	截距	-.805	.168	22.879	1	.000			
	[sex=1]	.458	.105	19.148	1	.000	1.581	1.288	1.941
	[sex=2]	0 <sup>b</sup>		.	0	.			
	[degree=0]	-.198	.228	.760	1	.383	.820	.525	1.281
	[degree=1]	.387	.175	4.913	1	.027	1.473	1.046	2.074
	[degree=2]	.431	.253	2.914	1	.088	1.539	.938	2.525
	[degree=3]	.424	.195	4.745	1	.029	1.529	1.043	2.239
	[degree=4]	0 <sup>b</sup>		.	0	.			
Perot	截距	-2.188	.264	68.527	1	.000			
	[sex=1]	.760	.140	29.319	1	.000	2.139	1.624	2.816
	[sex=2]	0 <sup>b</sup>		.	0	.			
	[degree=0]	-.502	.393	1.627	1	.202	.605	.280	1.309
	[degree=1]	.833	.267	9.709	1	.002	2.299	1.362	3.882
	[degree=2]	1.052	.346	9.263	1	.002	2.864	1.454	5.640
	[degree=3]	.804	.291	7.608	1	.006	2.233	1.262	3.953
	[degree=4]	0 <sup>b</sup>		.	0	.			

a. 参考类别是: Clinton。

b. 因为此参数冗余, 所以将其设为零。

表 11-39 所示是以克林顿为参考类模型中各参数及其检验结果。以上结论没有考虑性别和学历之间的交互作用。下面进行这方面的研究。其他选项与前面相同, 在【模型】选项中选择【全因子】模型, 在【统计量】选项中选择【似然比检验】, 其输出结果见表 11-40。

表 11-40 有交互项的似然比检验结果

由表 11-40 可见, 当交互项 sex\*degree 从方程式中剔除后,  $-2\ln$  的变化值的 Sig. 值很小, 大于 0.05, 也就是说, “把它们剔除出模型时并没有改变模型的拟合程度”。因此采用表 11-39 中的参数进行进一步分析。

4) 计算预测概率和预期频数

根据 Logistic 模型, 可以计算一个选民投票给某个候选人的可能性, 如具有学士学位的男性选民投票给各候选人的可能性。

估计每个分类的概率的公式为

效应	模型拟合标准	似然比检验		
	简化后的模型的-2 倍对数似然值	卡方	df	显著水平
截距	97.227 <sup>a</sup>	.000	0	.
sex	97.227 <sup>a</sup>	.000	0	.
degree	97.227 <sup>a</sup>	.000	0	.
sex * degree	103.601	6.374	8	.605

卡方统计量是最终模型与简化后模型之间在 -2 倍对数似然值中的差值。通过从最终模型中省略效应而形成简化后的模型。零假设就是该效应的所有参数均为 0。

$$p(\text{group}_i) = \frac{\exp(g_i)}{\sum_{k=1}^j \exp(g_k)}$$

式中,  $g_i$  是以最后一类作参考类, 第  $i$  类与参考类因变量值之比的概率的自然对数。这里很简单, 可以写出  $g_1/g_2$  的表达式计算其值, 根据 3 个候选人的被投票概率之和为 1, 列出联立方程得出解。

首先估算 3 个 Logit 模型的值, 根据表 11-35 的统计量可以分别计算出

$$\begin{cases} \ln p(\text{布什} / \text{克林顿}) = -0.805 + 0.458 + 0.424 = 0.077 & (g_1 \text{ 的值}) \\ \ln p(\text{帕洛特} / \text{克林顿}) = -2.188 + 0.760 + 0.804 = -0.624 & (g_2 \text{ 的值}) \\ p(\text{布什}) + p(\text{帕洛特}) + p(\text{克林顿}) = 1 \end{cases}$$

解联立方程得到具有学士学位的男性选民, 对每一位候选人投票的可能性。

$$\begin{aligned} p(\text{布什}) &= \frac{1.081}{1 + 1.081 + 0.535} = 0.413 \\ p(\text{帕洛特}) &= \frac{0.535}{1 + 1.081 + 0.535} = 0.205 \\ p(\text{克林顿}) &= \frac{1}{1 + 1.081 + 0.535} = 0.382 \end{aligned}$$

数据中有 160 名男性选民具有学士学位, 由此可以判断其中 66 人会投票给布什, 33 人会投给帕洛特, 61 人会投给克林顿。

5) 各类人实际投票与预测结果的比较

可以在【多项 Logistic 回归: 模型】对话框中将 sex、degree 设置成主效应, 在【多项 Logistic 回归: 统计量】对话框中选择【单元格概率】, 选择【分类表】。运行结果见表 11-41 和表 11-42。

表 11-41 所示是按性别、学历分组的实际和预测的单元格频数, 即百分比。表 11-42 所示是模型实际预测的正确率的分类统计表, 在实际投给布什选票的 661 人中有 251 人, 大约 38% 布什的支持者被模型正确地分类; 没有一个帕洛特的支持者被正确地分类; 大约 75% 的克林顿的支持者被模型正确地分类。总体来说, 被正确分类的占近 50%。这说明模型对数据的分类效果不佳。当按因变量分组的观测在几组中的数量差别较大时, 无论模型拟合有多好, 根据统计量预测的结果, 总是会把更多的观测分入包含大数据量的组中。

表 11-41 中的 Pearson 残差实际上是标准化残差  $Z_{ij}$ , 它的计算公式为

$$Z_{ij} = \frac{\alpha_{ij} - E_{ij}}{\sqrt{n_i(1 - \hat{p}_{ij})\hat{p}_{ij}}}$$

式中,  $\alpha_{ij}$  是实际观察值;  $E_{ij}$  是通过预测概率  $\hat{p}_{ij}$  计算得到的理论上期望出现的频数;  $n_i$  是每个亚群的合计频数。

以表 11-41 中第一个格子中的数据为例, 该亚群的  $n_i = 27 + 6 + 50 = 83$ , 投票给布什的理论期望频数为 27.902, 其预测概率  $\hat{p}_{ij} = 27.902/83 = 0.336169$ 。因此, 根据 Pearson 残差的计算公式

可以得到 Pearson 残差值为  $\frac{27 - 27.902}{\sqrt{83 \times 0.336169(1 - 0.336169)}} = -0.210$ , 其余可类推。

表 11-41 观测值与预测值比较

观察频率和预测频率							
RS HIGHEST DEGREE	RESPONDENTS SEX	VOTE FOR CLINTON, BUSH, PEROT	频率			百分比	
			观察值	预测值	Pearson 残差	观察值	预测值
lt high school	male	Bush	27	27.902	-.210	32.5%	33.6%
		Perot	6	6.985	-.389	7.2%	8.4%
		Clinton	50	48.114	.419	60.2%	58.0%
	female	Bush	28	27.098	.201	26.4%	25.6%
		Perot	6	5.015	.450	5.7%	4.7%
		Clinton	72	73.886	-.399	67.9%	69.7%
high school	male	Bush	158	162.701	-.476	39.0%	40.2%
		Perot	89	86.103	.352	22.0%	21.3%
		Clinton	158	156.197	.184	39.0%	38.6%
	female	Bush	191	186.299	.425	35.2%	34.4%
		Perot	70	72.897	-.365	12.9%	13.4%
		Clinton	281	282.803	-.155	51.8%	52.2%
junior college	male	Bush	22	21.965	.010	39.3%	39.2%
		Perot	17	13.856	.974	30.4%	24.7%
		Clinton	17	20.179	-.885	30.4%	36.0%
	female	Bush	26	26.035	-.009	34.2%	34.3%
		Perot	9	12.144	-.984	11.8%	16.0%
		Clinton	41	37.821	.729	53.9%	49.8%
bachelor	male	Bush	71	66.108	.785	44.4%	41.3%
		Perot	27	32.743	-1.125	16.9%	20.5%
		Clinton	62	61.149	.138	38.8%	38.2%
	female	Bush	75	79.892	-.681	33.2%	35.4%
		Perot	35	29.257	1.138	15.5%	12.9%
		Clinton	116	116.851	-.113	51.3%	51.7%
graduate degree	male	Bush	37	36.325	.140	37.0%	36.3%
		Perot	13	12.314	.209	13.0%	12.3%
		Clinton	50	51.361	-.272	50.0%	51.4%
	female	Bush	26	26.675	-.155	28.0%	28.7%
		Perot	6	6.686	-.275	6.5%	7.2%
		Clinton	61	59.639	.294	65.6%	64.1%

百分比基于各个子总体中的总观察频率。

6) 检验拟合的优劣

在【多项 Logistic 回归：统计量】对话框中选择【拟合度】，得出表 11-43。由于 Pearson、偏差(翻译有误，应为 Deviance)卡方统计量的 Sig 值全部大于 0.05，从而判断出模型对数据拟合较好。只有当协变量可以看作有序分类变量，且各分组单元都有大量的观测时，才能使用模型的拟合度统计量；如果协变量各单元格分组中观测数差别很大，拟合度统计量不适用。

表 11-42 分类表

观察值	预测值			
	Bush	Perot	Clinton	百分比校正
Bush	251	0	410	38.0%
Perot	133	0	145	0.0%
Clinton	237	0	671	73.9%
总百分比	33.6%	0.0%	66.4%	49.9%

表 11-43 拟合度统计量

拟合优度			
	卡方	df	显著水平
Pearson	6.327	8	.611
偏差	6.374	8	.605

Goodness-of-Fit			
	Chi-Square	df	Sig.
Pearson	6.327	8	.611
Deviance	6.374	8	.605

11.5 有序变量 Logistic 回归

11.5.1 有序变量 Logistic 回归的概念

在前面讨论的 Logistic 回归中，因变量无论是二分类的，还是多分类的，它们都是名义测度的变量。名义变量的各类之间是按其属性的不同来划分的，类与类之间只是名义上的区别，没有本质的高低、大小或轻重之分，但在实际的工作中，会经常遇到多分类变量的各类之间在其属性上还会有轻重、大小、高低或程度的不同之分。例如，运动员等级变量，它的取值可以分为 5 种：1—国际健将、2—国家健将、3—国家一级、4—国家二级、5—其他，显然，这些

不同的级别是按运动员水平由高到低进行排列的，但相互之间的差距却不一定是等间隔的。又如，患者在用药物进行治疗时，对不同药物剂量的反应可以分为无、轻微、适度或剧烈。轻微反应和适度反应之间的差别取决于感觉，很难或不可能量化。另外，轻微反应和适度反应之间的差别可能比适度反应和剧烈反应之间的差别更大或更小。尽管如此，在反应的程度上还是有轻重之分的。这种按属性的不同程度进行分类得到的资料，称为有序资料，而描述有序资料的变量就称为有序变量。

对有序变量进行预测时，可以用有序变量 Logistic 回归。

1. 有序变量 Logistic 回归模型

SPSS 中的有序变量 Logistic 回归是以 Mc Cullagh(1980, 1998)提出的方法为基础的，其数学模型为

$$\eta_{ij}[\pi_{ij}(Y \leq j)] = \frac{\alpha_j - (\beta_1 X_{i1} + \cdots + \beta_p X_{ip})}{\sigma_i} \quad j=1,2,\cdots,J-1$$

式中， $i$  ( $i=1,2,\cdots,m$ ) 表示分组数(自变量向量的行数)； $j$  ( $j=1,2,\cdots,J$ ) 表示因变量  $Y$  的分类数； $k$  ( $k=1,2,\cdots,p$ ) 表示自变量 ( $X_1,\cdots,X_p$ ) 的个数； $\alpha_j$  为常数项 ( $j=1,2,\cdots,J-1$ )； $\beta_k$  为回归系数 ( $k=1,2,\cdots,p$ )； $\sigma_i$  为尺度参数(默认值为 1)； $\pi_{ij}(Y \leq j) = \pi_{i1} + \cdots + \pi_{ij}$  为因变量  $Y \leq j$  的累积概率； $\eta_{ij}[\pi_{ij}(Y \leq j)]$  为关于累积概率  $\pi_{ij}(Y \leq j)$  的链接函数。

链接函数是累积概率的转换形式，可用于模型估计。在 SPSS 中，主要可从以下 5 种链接函数中选择：

- Cauchit 链接函数  $\tan\{\pi_i(Y=j)[\pi_i(Y \leq j)-0.5]\}$ ；
- 补对数对数链接函数  $\ln\{-\ln[1-\pi_{ij}(Y \leq j)]\}$ ；
- Logit 链接函数  $\ln\frac{\pi_{ij}(Y \leq j)}{1-\pi_{ij}(Y \leq j)}$ ；由此形成的模型称为累加 Logit 模型，也称为比例优势模型；该函数是 SPSS 中默认的链接函数；
- 负对数对数链接函数  $\ln\{-\ln[\pi_{ij}(Y \leq j)]\}$ ；
- 概率链接函数  $\Phi^{-1}[\pi_{ij}(Y \leq j)]$ 。 $\Phi^{-1}(\cdot)$  为标准正态分布分位数。

这些链接函数的适用条件归纳见表 11-44。

表 11-44 链接函数的适用条件

函 数	形 式	典 型 应 用
Logit 链接函数	$\ln [\xi / (1-\xi)]$	均匀分布类别
补对数对数链接函数	$\ln [-\ln (1-\xi)]$	类别越高可能性越大
负对数对数链接函数	$-\ln [-\ln (\xi)]$	类别越低可能性越大
概率链接函数	$\Phi^{-1}(\xi)$	潜在变量为正态分布
Cauchit 链接函数	$\tan [\pi(\xi-0.5)]$	潜在变量有许多个极值

SPSS 系统默认的链接函数之所以要设为 Logit 链接函数，是因为比例优势模型是有序变量 Logistic 回归中最常用的模型。该模型为

$$\ln \frac{\pi_{ij}(Y \leq j)}{1-\pi_{ij}(Y \leq j)} = \ln \frac{\pi_{i1} + \cdots + \pi_{ij}}{\pi_{i(j+1)} + \cdots + \pi_{iJ}} = \alpha_j - (\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) \quad j=1,2,\cdots,J-1$$



由此可得累加 Logit 的  $J-1$  个预测概率模型为

$$\pi_{ij}(Y \leq j) = \frac{\exp[\alpha_j - (\beta_1 X_{i1} + \beta_p X_{ip})]}{1 + \exp[\alpha_j - (\beta_1 X_{i1} + \cdots + \beta_p X_{ip})]} \quad j=1, 2, \dots, J-1$$

累积概率具有以下两个性质:

- ①  $\pi(Y \leq 1) \leq \pi(Y \leq 2) \leq \cdots \leq \pi(Y \leq J)$ ;
- ②  $\pi(Y \leq J) = 1$ 。

## 2. 模型对数据的要求

(1) 数据。因变量须为有序变量,可以是数值或字符串。通过对因变量的值进行升序排序来确定排列顺序,最低值定义第一个类别。因子变量须为分类变量,协变量必须为连续型数值变量。需要注意的是:如果使用多个连续型协变量,则很容易使创建的单元概率表非常大。

(2) 假设。只允许使用一个因变量,且必须指定该因变量。另外,对于多个自变量值的各个不同模式,假设该变量是独立的多分类变量。

由于标准的 Logistic 回归对于名义因变量使用相似的模型,因而有序变量 Logistic 回归模型参数的意义、解释及模型的假设检验、模型的拟合优度评价方法与二项 Logistic 回归相似,这里不再赘述。

### 11.5.2 有序变量 Logistic 回归过程

- (1) 按【分析→回归→有序】顺序打开如图 11-56 所示的【Ordinal 回归】对话框。
- (2) 在左侧的源变量框中选择一个有序变量作为因变量进入【因变量】框中。
- (3) 在左侧的源变量框中选择一个或多个分类变量进入【因子】框中。

**注意:** 这里哑变量编码以数字较大者作为参照类别。

(4) 在左侧的源变量框中选择一个或多个连续型变量或 0、1 二分类变量进入【协变量】框中。

(5) 单击【选项】按钮,弹出【Ordinal 回归: 选项】对话框,见图 11-57。在此对话框中可以调整迭代估计算法中所使用的参数,选择参数估计值的置信度并选择链接函数。



图 11-56 【Ordinal 回归】对话框



图 11-57 【Ordinal 回归: 选项】对话框

- ① 【迭代】栏。设置迭代停止的判定标准。
  - 【最大迭代】。设置最大迭代数。它必须为小于等于 100 的正整数。系统默认值为 100。如果指定为 0, 则过程会返回初始估计值。
  - 【最大步骤对分】(最大逐步二分法)框。输入使用逐步二分法的最大步数。系统默认值为 5。

- **【对数似然性收敛性】**框。可以给定一个正数来设置对数似然比收敛值。当回归过程中的对数似然比大于此值时,迭代过程将停止。系统默认值为 0。
- **【参数收敛性】**框。设置收敛参数。在模型拟合过程中,如果绝对变化值或相对变化值大于等于此值时,迭代过程将停止。系统默认值为 0.000001。

② **【置信区间】**框。指定一个大于等于 0 且小于 100 的值。系统默认值为 95。

③ **【Delta】**框。输入小于 1 的非负值,此值会出现在交叉表的空单元中。系统默认值为 0。这将有助于稳定算法、阻止估计偏差。

④ **【奇异性容许误差】**下拉列表。选择检验单一性的容忍度值。系统默认值为 0.00000001。

⑤ **【链接】**下拉列表。选择一种链接方式。系统默认为 Logit。

#### (6) 输出结果设置

单击**【输出】**按钮,弹出**【Ordinal 回归: 输出】**对话框,见图 11-58。使用该对话框可以生成在浏览器中显示的表格,并将变量保存到当前数据文件中。

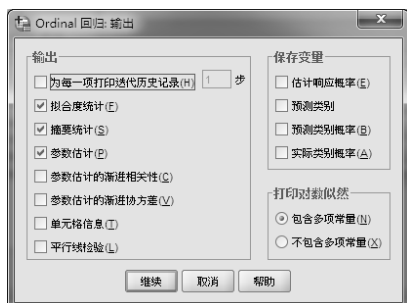


图 11-58 **【Ordinal 回归: 输出】**对话框

#### ① **【输出】**栏。

- **【为每一项打印迭代历史记录】**的后框中,设置输出迭代过程的步距。系统默认值为 1。它为所指定的打印对数似然估计和参数估计值。始终打印第一个和最后一个迭代中的对数似然估计和参数估计值。
- **【拟合优度统计】**。根据在变量列表中指定的分类来计算 Pearson 卡方和似然比卡方统计量,并在输出窗中输出。
- **【摘要统计】**。在输出窗中输出有 Cox 和 Snell、Nagelkerke 和 McFadden  $R^2$  统计量信息的统计表。
- **【参数估计】**。在输出窗中输出有参数估计值、标准误和置信区间信息的统计表。
- **【参数估计的渐近相关性】**。在输出窗中输出参数估计的相关系数矩阵。
- **【参数估计的渐近协方差】**。在输出窗中输出参数估计的协方差矩阵。
- **【单元格信息】**。在输出窗中为因变量各类输出因子变量与协变量各类组合中的观察频数、期望频数及 Pearson 残差等的信息表。
- **【平行线检验】**。作在多个因变量水平上位置参数均相等的假设检验。该检验仅适用于比例优势模型。

#### ② **【保存变量】**栏。

在当前的工作文件中以新变量的方式保存以下选择项的信息:

- **【估计响应概率】**。因变量每一个类别的每一个格子的预测概率。
- **【预测类别】**。每一个格子的预测类别。
- **【预测类别概率】**。每一个格子的预测类别对应的预测概率。
- **【实际类别概率】**。将每一个格子的实际类别对应的预测概率。

#### ③ **【打印对数似然】**栏。可控制对数似然估计的显示。

- **【包含多项式常量】**。输出包括常数项对数似然估计值。
- **【不包含多项式常量】**。输出不包括常数项对数似然估计值。

(7) 指定分析的位置模型

在主对话框中单击【位置】按钮，弹出【有序回归：位置模型】对话框，见图 11-59。

- ①【指定模型】栏。可以选择【主效应】选项，也可以用【设定】选项来自定义模型。
- 【主效应】。在模型中只包括在主对话框的【因子和协变量】框中所选定变量的主效应，不包括因子和协变量的交互效应。
  - 【设定】。用户可自行设定模型中包括的主效应和交互效应。
- ②【因子/协变量】框。列出主对话框中选定的因子与协变量。

【构建项】栏的【类型】下拉列表有以下可选项：

- A.【主效应】。为每个选定的变量创建主效应项。
- B.【交互】。为所有选定变量创建最高阶交互项。（它是系统默认选项。）
- C.【所有二阶】。为所选定变量创建所有可能的二阶交互项。
- D.【所有三阶】。为所选定变量创建所有可能的三阶交互项。
- E.【所有四阶】。为所选定变量创建所有可能的四阶交互项。
- F.【所有五阶】。为所选定变量创建所有可能的五阶交互项。

(8) 指定分析

单击【度量】按钮，弹出【Ordinal 回归：度量模型】对话框，见图 11-60 的尺度模型。指定分析模型设定方法同前。



图 11-59 【有序回归：位置】对话框



图 11-60 【Ordinal 回归：度量】对话框

11.5.3 有序变量的 Logistic 回归分析实例

【例 8】某研究者分别于 1985 年、1995 年、2005 年调查了已婚及未婚的 30 岁左右成年人的幸福感情况，调查结果见表 11-45。分析不同年份、不同婚姻状况的被调查者的幸福感有何不同。

表 11-45 不同年份、不同婚姻状况的幸福感

年份	婚姻状况	幸福感程度		
		不太幸福	比较幸福	十分幸福
1985	已婚	214	869	237
	未婚	93	773	551
1995	已婚	80	211	65
	未婚	76	473	453
2005	已婚	98	327	130
	未婚	46	367	312

1) 操作步骤

(1) 在 SPSS 中建立 4 个变量用来存放表 11-45 中的数据。

① 年份变量，数值型，值标签：3-1985、2-1995、3-2005，名义测度，用来存放表中的年份数据。

② 婚姻状况变量，数值型，值标签：1—已婚、0—未婚，名义测度，用来存放表中的婚姻状况数据。

③ 幸福感程度变量，数值型，值标签：1—不太幸福、2—比较幸福、3—十分幸福，有序测度，用来存放表中的幸福感程度数据。

④ 频数变量，数值型，尺度测度，用来存放表中的调查结果数据。

由此建立的数据文件为 data11-06。

(2) 按【数据→加权个案】顺序打开【加权个案】对话框，选择【加权个案】选项，将“频数”移入【频率(应为频数)变量】框中，单击【确定】按钮，完成加权处理工作。

(3) 按【分析→回归→有序】顺序打开【Ordinal 回归】对话框。

在左侧的源变量框中选择“幸福感程度”进入【因变量】框中，选择“年份”进入【因子】框中，选择“婚姻状况”进入【协变量】框中。

(4) 单击【输出】按钮，弹出【Ordinal 回归：输出】对话框，选择【拟合度统计】、【摘要统计】、【参数估计】、【单元格信息】、【平行线检验】选项。单击【继续】按钮返回【Ordinal 回归】对话框。

(5) 单击【确定】按钮提交运算，在输出窗中得到表 11-46～表 11-52 所示的输出结果。

2) 结果分析

表 11-46 所示为因变量及因子(自变量)的每个类别的频数及构成比情况。有 56.2%的被试者认为比较幸福。在被调查的人员中，1985 年调查的人数最多，占总被调查人数的百分比为 50.9%。

表 11-47 所示为模型整体拟合信息表。从表中可见， $p = 0.000$ ，小于 0.05，表明最终的模型有统计上的显著性意义。

表 11-48 所示为模型的拟合优度检验。从表中可见，Pearson 卡方、Deviance 卡方统计量的  $p$  值均小于 0.05，说明模型拟合较差。

表 11-46 个案处理摘要

案例处理摘要			
		N	边际百分比
幸福感	不太幸福	607	11.3%
	比较幸福	3020	56.2%
	十分幸福	1748	32.5%
年份	2005	1280	23.8%
	1995	1358	25.3%
	1985	2737	50.9%
有效		5375	100.0%
缺失		0	
合计		5375	

表 11-47 模型拟合信息

表 11-48 模型的拟合优度

模型拟合信息				
模型	-2 对数似然值	卡方	df	显著性
仅截距	486.570			
最终	102.905	383.665	3	.000

联接函数：Logit。

拟合度			
	卡方	df	显著性
Pearson	25.091	7	.001
偏差	24.634	7	.001

联接函数：Logit。

表 11-49 伪 R 方

伪 R 方	
Cox and Snell	.069
Nagelkerke	.081
McFadden	.038

联接函数：Logit。

表 11-49 所示是 3 种方法的伪 R 方值，这些值相对较小，都不足 9%，结合表 11-46～表 11-48 的分析结果可知模型不够理想，可以考虑用其他模型来拟合。

表 11-50 所示为参数估计及其检验结果值。婚姻状况的估计值为 1.07，优势比= $\exp(1.07)=2.94$ ，说明已婚者的幸福感高于未婚者，约为其 3 倍；

由于与 1985 年相比, 2005 年的估计值为 0.141, 1995 年的估计值为 0.084, 都为正数, 也都很小, 说明 1995 年和 2005 年相对于 1985 年而言, 幸福感均有所提高, 但优势并不明显。

表 11-50 参数估计及其检验结果

参数估计值							95% 置信区间	
		估计	标准误	Wald	df	显著性	下限	上限
阈值	[幸福感 = 1]	-1.492	.054	750.258	1	.000	-1.598	-1.385
	[幸福感 = 2]	1.468	.054	740.099	1	.000	1.362	1.573
位置	婚姻状况	1.077	.058	341.008	1	.000	.963	1.192
	[年份=1]	.141	.067	4.428	1	.035	.010	.272
	[年份=2]	.084	.067	1.601	1	.206	-.046	.215
	[年份=3]	0 <sup>a</sup>	.	.	0	.	.	.

联接函数：Logit。

a. 因为该参数为冗余的，所以将其置为零。

根据第 11.5.1 节中的公式, 可分别列出不同亚群间的两个累加预测概率的 Logit 模型。  
以年份 1 (2005 年) 为例, 婚姻状态为未婚的对于幸福感程度的两个累加预测概率的 Logit 模型为

$$\hat{p}(\text{幸福感程度} \leq 1) = \frac{\exp(-1.492 - 1.077 \times 0 - 0.141)}{1 + \exp(-1.492 - 1.077 \times 0 - 0.141)} = 0.16348$$
$$\hat{p}(\text{幸福感程度} \leq 2) = \frac{\exp(1.468 - 1.077 \times 0 - 0.141)}{1 + \exp(1.468 - 1.077 \times 0 - 0.141)} = 0.79029$$

因此, 该类人群中, 幸福感程度为不太幸福的概率值为 0.16348, 比较幸福的概率值为 0.79029 - 0.16348 = 0.62681, 十分幸福的概率为 1 - 0.79029 = 0.20971。

由表 11-51 中可见, 该类人群的人数为 98+327+130=555。基于上述各类概率可以计算得到这类人群不太幸福、比较幸福、十分幸福的期望值分别为 0.16348 × 555 = 90.732、0.62681 × 555 = 347.879、0.20971 × 555 = 116.389, 这与表 11-51 中的结果相同。

与上述算法顺序反向进行操作, 则根据表 11-51 中的结果可以反过来推算出各类人群的幸福感的 Logit 模型的预测概率。以年份 1 (2005 年) 已婚者人群的幸福感的程度的数据为例, 具体做法如下:

先求该人群参与调查的总人数 46+367+312=725, 再用该类人群的各期望值与 725 之比, 也就是 45.231/725、362.213/725、317.556/725 得到这类人群的幸福感的程度的 Logit 模型的预测概率, 它们分别为 0.062387、0.499604、0.438009, 这与用下面该类的 2 个累加预测概率的 Logit 模型计算得到的结果一致。读者可根据前面的做法加以验证。

表 11-51 单元格信息

频率			幸福感		
年份	婚姻状况		不太幸福	比较幸福	十分幸福
2005	未婚	观察值	98	327	130
		期望值	90.732	347.879	116.389
		Pearson 残差	.834	-1.832	1.419
	已婚	观察值	46	367	312
		期望值	45.231	362.213	317.556
		Pearson 残差	.118	.356	-4.16
1995	未婚	观察值	80	211	65
		期望值	61.021	223.622	71.357
		Pearson 残差	2.669	-1.384	-.842
	已婚	观察值	76	473	453
		期望值	65.930	511.160	424.910
		Pearson 残差	1.283	-2.412	1.796
1985	未婚	观察值	214	869	237
		期望值	242.478	830.299	247.223
		Pearson 残差	-2.024	2.205	-.721
	已婚	观察值	93	773	551
		期望值	100.841	744.192	571.968
		Pearson 残差	-.810	1.533	-1.135

联接函数：Logit。

该类的 2 个累加预测概率的 Logit 模型为

$$\hat{p}(\text{幸福感程度} \leq 1) = \frac{\exp(-1.492 - 1.077 \times 1 - 0.141)}{1 + \exp(-1.492 - 1.077 \times 1 - 0.141)} = 0.062387$$
$$\hat{p}(\text{幸福感程度} \leq 2) = \frac{\exp(1.468 - 1.077 \times 1 - 0.141)}{1 + \exp(1.468 - 1.077 \times 1 - 0.141)} = 0.561991$$

Pearson 残差值的计算方法与【例 7】相同，从略。

表 11-52 平行线检验<sup>a</sup>

模型	-2 对数似然值	卡方	df	显著性
零假设	102.905			
广义	83.626	19.280	3	.000

零假设规定位置参数（斜率系数）在各响应类别中都是相同的。

联接函数：Logit。

表 11-52 显示了平行线检验的结果。由于  $p=0.000$ ，说明有充分的证据可拒绝多个因变量水平上位置参数均相等的假设，表明需改用其他连接函数来作有序变量的 Logistic 回归分析，如效果仍然不佳，则表明系数的确在随着分割点的不同而发生变化，可改用无序多分类的 Logistic 回归进行建模分析。

11.6 概率单位回归

11.6.1 概率单位回归的概念

1. 概率单位回归分析

概率单位回归在 SPSS 软件中属于专业统计分析过程，用来分析反应比例与刺激强度之间的关系。例如，研究一定数量的病人给药剂量与治愈的百分比之间的关系。

由于线性模型的某些限制，需要把可能分布在整个实数轴上的  $x$  值通过累计概率函数  $f$  变换成分布在  $(0,1)$  区间中的概率值，概率分布表达式为

$$p_i = f(\alpha + \beta x_i) = f(Z_i)$$

概率单位回归分析只考虑诸多累计概率函数中的两种。

(1) 标准正态累计概率函数：

$$p_i = F(Z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2} ds$$

式中， $p_i$  是事件发生的概率； $s$  是零均值单位方差的正态分布的随机变量。由于这个概率是标准正态分布函数曲线下  $-\infty \sim Z$  之间的面积，所以  $Z_i$  的值越大，事件就越可能发生。

(2) Logit 概率函数  $p_i = F(Z_i) = F(\alpha + \beta x_i) = \frac{1}{1 + e^{-Z_i}} = \frac{1}{e^{-(\alpha + \beta x_i)}}$ ，通过转换可以得到

$$\text{Ln} \frac{P_i}{1 - P_i} = Z_i = \alpha + \beta x_i$$

例如，可以设计一个试验，记录不同浓度杀虫剂杀死白蚁的数量。使用概率单位回归分析，就可以得出杀虫剂浓度与杀死白蚁数量的关系，据此判明什么样的杀虫剂浓度是最佳的(如可以杀死 95%以上的白蚁)。药学中，此方法常用于半数效量研究，即求完成 50%反应的刺激量。

再如，可以用来检测购买某类物品的人员比例与所提供的物品刺激数量之间的关系，在研究的数据具有相反的属性(如买与不买)，或者几组研究对象被作用于不同水平刺激条件而产生不同反应水平时才能应用概率单位回归分析。

2. 概率单位分析与 Logistic 分析的区别

概率单位模型实际上是由 Logit 模型和 Probit 模型组成的。因此，首先利用 Logit 和 (或) Probit 过程来转换响应比例，而不是直接使用“刺激”所产生的响应比例进行回归计算。

表 11-53 表明 Probit 和 Logit 的公式十分相似。因为 Logit 概率分布函数与正态分布密度函数近似，所以常用 Logit 模型来替代 Probit 模型。

3. 数据要求

- (1) 因变量中的每个数据应该是对某一水平刺激发生反应的数量。
- (2) 观测值应该是独立的，否则卡方检验和拟合优度检验是不适宜的。

表 11-53 概率分布函数值比较

Z	正态累计概率函数 $p_i(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2} ds$	Logit 概率函数 $p_i(Z) = \frac{1}{1 + e^{-Z_i}}$
-3.0	0.0013	0.0474
-2.0	0.0228	0.1192
-1.5	0.0668	0.1824
-1.0	0.1587	0.2689
-0.5	0.3085	0.3775
0.0	0.5000	0.5000
0.5	0.6915	0.6225
1.0	0.8413	0.7311
1.5	0.9332	0.8176
2.0	0.9772	0.8808
3.0	0.9987	0.9526

11.6.2 概率单位回归过程

- (1) 按【分析→回归→Probit】顺序打开如图 11-61 所示的对话框。
  - (2) 选择一个变量作为响应频数变量进入【响应频率】框中。这个变量中的每个数值是对实验刺激水平作出反应的观测值的数目总和。该变量的值不能为负数。
  - (3) 选择一个变量作为总观测变量进入【观测值汇总】框中。这个变量是用于某一刺激水平的观测值总数。该变量的值不能小于响应频数变量的值。
  - (4) 可选择一个因素变量进入【因子】框。单击【定义范围】按钮，在对话框中给出因素变量的最小值和最大值。
  - (5) 选择至少一个协变量进入【协变量】框中。协变量是不相同的试验刺激条件值。
- 协变量和 Probit(p) 之间不存在线性关系时，在【转换】下拉列表中选择转换模式，对协变量进行转换。3 个选项分别为：【无】，不进行转换 (默认)；【对数底为 10】，用以 10 为底的对数进行转换；【自然对数】，用以 e 为底的自然对数进行转换。至于是否进行转换或选择哪种转换，要选择不同的转换方法，经过几次运行概率单位回归过程，比较分析结果再确定，同时得出分析结论。
- (6) 在【模型】栏中确定一种算法。
    - ① 【概率】。用累积标准正态分布函数的反函数来转换响应比例。
    - ② 【Logit】。对响应比例应用自然对数转换。
  - (7) 单击【选项】按钮，打开如图 11-62 所示的对话框。
    - ① 【统计量】栏。输出统计量。
      - 【频率】 (应为频数)。输出每一个观测值与预测值的频数以及每一个观测值的残差。
      - 【相关中位数力】 (应为相对中位数潜力)。输出因素变量各水平间中位数比较的效应及 95% 的置信区间。如果模型中没有因子变量或具有多个协变量，则不可以用它。
      - 【平行检验】。平行检验的假设是因素变量各分组回归方程具有相同的斜率。
      - 【信仰置信区间】 (应为置信信赖区间)。如果选择了因素变量，可选此项。在【异质因子使用的显著性水平】框中输入一个显著性水平值，将对因素变量的每个水平显示从 0.01~

0.99 反应比例所需的刺激强度的置信区间。当拟合优度值小于设定值时, Probit 用非齐性修正方法计算置信区间。选择了协变量, 就不适用置信区间与半数有效量的计算。



图 11-61 【Probit 分析】对话框

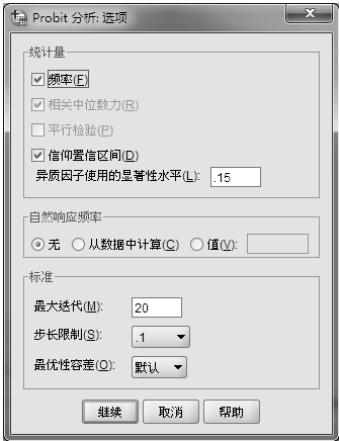


图 11-62 【Probit 分析: 选项】对话框

②【自然响应频率】栏。设置是否计算自然响应率。没有刺激条件下的响应称为自然响应。例如, 如果试验对象生命较短, 在试验过程中会发生一些自然死亡, 这时就需要调整观测比例以反映真实的“刺激”条件所产生的响应。

- 【无】。不计算自然响应率。
- 【从数据中计算】。根据提供的数据计算刺激强度为零的响应观测值。
- 【值】。输入小于 1 的已知自然响应频率。例如, 自然响应率是 12% 时, 输入“0.12”。

③【标准】栏。设置控制迭代停止的判定标准。

- 【最大迭代】框。输入控制迭代停止的最大迭代步数。
- 【步长限制】下拉列表。选择参数向量所容许的最大变化量。
- 【最优性容差】下拉列表。设定损失函数的精确值。

(8) 单击【确定】按钮进行统计分析。

### 11.6.3 概率单位回归分析实例

【例 9】数据文件 data11-07 记录了不同杀虫剂、不同浓度、不同杀虫效果的数据。变量包括: died 各组白蚁死亡数、total 各组白蚁总数、dose 杀虫剂剂量、agent 杀虫剂类别。使用这 4 个变量求各种杀虫剂的半数致死量。

#### 1) 操作步骤

- (1) 按【分析→回归→Probit】顺序打开对话框。
- (2) 选择“died”作为响应变量送入【响应频率】框; 选择“total”作为总观测变量送入【观测值汇总】框中。选择剂量变量 dose 送入【协变量】框中。
- (3) 选择“agent”作为因素变量送入【因子】框中, 单击【定义范围】按钮, 打开对话框, 在【最小值】后输入“1”, 在【最大值】后输入“3”。
- (4) 在【转换】下拉列表中选择【对数底为 10】, 作为第一次运行该分析过程的选择。
- (5) 在【Probit 分析: 选项】对话框中选择【平行检验】, 其他参数选项均为默认值。
- (6) 单击【确定】按钮进行统计分析。输出结果见表 11-54~表 11-58 和图 11-63。



2) 结果分析

表 11-54 给出了数据的基本情况。共有 15 个合法观测值，没有观测值被剔除，3 种杀虫剂 deguelin 鱼藤素、rotenone 鱼藤酮、mixture 混合物的观测值数均为 5 个。

在表 11-55 中，第一个表说明进行 15 步迭代后，找到了最佳结果；

表 11-54 数据基本统计

数据信息	
	个案数
有效	15
已拒绝	0
超出范围 <sup>a</sup>	0
缺失	0
不能执行对数转换	0
响应数 > 主体数	0
控制组	0
agent deguelin	5
rotenone	5
mixture	5

a. 由于超出组值范围，个案被拒绝。

表 11-55 模型参数

收敛信息

	迭代数	找到最优解
PROBIT	15	是

参数估计值

		估计	标准误	z	Sig.	95% 置信区间	
						下限	上限
PROBIT <sup>a</sup>	dose	4.006	.274	14.640	.000	3.469	4.542
	截距 <sup>b</sup>						
	deguelin	-2.743	.214	-12.800	.000	-2.958	-2.529
	rotenone	-4.492	.366	-12.274	.000	-4.858	-4.126
	mixture	-2.741	.214	-12.809	.000	-2.955	-2.527

a. PROBIT 模型:  $\text{PROBIT}(p) = \text{截距} + \text{BX}$  (协变量 X 使用底数为 10.000 的对数来转换。)

b. 对应于分组变量 agent。

卡方检验

	卡方	df <sup>b</sup>	Sig.
PROBIT Pearson 拟合度检验	9.374	11	.587 <sup>a</sup>
平行检验	1.664	2	.435

a. 由于显著性水平大于 .150，因此在置信限度的计算中未使用异质因子。

b. 基于单个个案的统计量与基于分类汇总个案的统计量不同。

第二个表是参数估计表。给出了方程形式，3 种杀虫剂效果的模型分别为：

- 杀虫剂 deguelin 鱼藤素的方程： $\text{Probit}(p) = -2.743 + 4.006\lg(\text{dose})$ ；
- 杀虫剂 rotenone 鱼藤酮的方程： $\text{Probit}(p) = -4.492 + 4.006\lg(\text{dose})$ ；
- 杀虫剂 mixture 混合物的方程： $\text{Probit}(p) = -2.741 + 4.006\lg(\text{dose})$ 。

第三个表中，Pearson 拟合优度卡方检验的显著水平为 0.587，大于 0.05，拟合良好。

如果 Pearson 卡方显著水平值较小，或是因为药剂量与  $\text{Probit}(p)$  之间没有存在线性关系，或虽为线性，但观测值在直线周围的分布不均匀。

由于平行检验的  $p$  值为 0.435，大于 0.05，不足以拒绝零假设(不排除在更多样本或另一个检验方法时拒绝零假设)，即 3 种杀虫剂方程式直线相互平行。

表 11-56 所示为 3 种杀虫剂各剂量 dose 的致死率 Probit 及 95%置信区间上下限。表中可查 3 种杀虫剂的半数致死量，即  $\text{Probit} = 0.5$  时的剂量的估计值分别为 4.840、13.229、4.833。

表 11-57 所示是按因素变量分组所得的观测值与期望值数据，包括杀虫剂类别 agent 为分组变量、dose 为剂量、主体数(应为被试对象的数量)、观测的响应频数、期望的响应频数、残差、概率。

表 11-58 所示为各组中位数效应比值。杀虫剂 deguelin 的中位数为 4.84(表 11-56 中概率 0.5 对应的估计值)，rotenone 的中位数为 13.229，因此，杀虫剂 deguelin 对 rotenone 的中位数比值为  $4.84/13.22=0.366$ ，mixture 对 rotenone 的比值为 1.001，rotenone 对 mixture 的比值为 2.737。

图 11-63 所示为 3 种杀虫剂剂量取对数与概率值的散点图。从图中可以看出，概率值与不同刺激剂量呈现较为明显的线性关系，说明取“对数底为 10”的选项进行转换是比较合适的。如果散点图没有呈现线性关系，那么还需要进行其他方法的转换，或各种转换各做一次，比较其结果。一定要确保转换后数据的线性关系。

表 11-56 3 种杀虫剂各剂量致死率与 95%的置信区间

agent	剂量	概率	dose 的 95% 置信上限			log(dose) 的 95% 置信上限		
			估计	下限	上限	估计	下限	上限
PROBIT	010	1.271	.389	1.538	.164	-.181	.187	
	020	1.487	1.192	1.772	.172	.079	.248	
	030	1.942	1.335	1.838	.216	.125	.287	
	040	1.781	1.482	2.074	.248	.182	.317	
	050	1.881	1.555	2.182	.274	.182	.341	
	060	1.861	1.648	2.286	.287	.217	.381	
	070	2.012	1.735	2.385	.318	.239	.379	
	080	2.158	1.818	2.488	.354	.259	.388	
	090	2.242	1.883	2.572	.388	.277	.418	
	100	2.317	1.988	2.654	.385	.284	.424	
	110	2.881	2.289	3.023	.428	.382	.458	
	120	2.884	2.881	3.351	.493	.415	.528	
	130	3.281	2.888	3.678	.518	.481	.581	
	140	3.581	3.171	3.882	.554	.581	.681	
	150	3.871	3.454	4.512	.588	.588	.835	
roteneone	010	3.473	2.589	4.381	.541	.413	.642	
	020	4.063	3.084	5.045	.609	.490	.703	
	030	4.487	3.463	5.519	.652	.538	.742	
	040	4.836	3.768	5.806	.684	.576	.771	
	050	5.139	4.036	6.241	.711	.606	.795	
	060	5.412	4.279	6.542	.733	.631	.816	
	070	5.684	4.504	6.818	.753	.654	.834	
	080	5.889	4.714	7.075	.771	.673	.850	
	090	6.121	4.914	7.318	.787	.691	.864	
	100	6.333	5.106	7.549	.802	.708	.878	
	110	7.291	5.977	8.592	.863	.776	.934	
	120	8.155	6.770	9.529	.911	.831	.979	
	130	8.977	7.528	10.420	.953	.877	1.018	
	140	9.786	8.277	11.298	.991	.918	1.053	
	150	10.600	9.033	12.184	1.025	.956	1.086	
mixture	010	1.268	.997	1.538	.103	-.001	.188	
	020	1.484	1.191	1.789	.172	.078	.248	
	030	1.640	1.333	1.935	.215	.125	.287	
	040	1.787	1.451	2.071	.247	.162	.316	
	050	1.878	1.554	2.188	.274	.191	.340	
	060	1.977	1.647	2.294	.296	.217	.361	
	070	2.089	1.733	2.391	.316	.239	.379	
	080	2.155	1.814	2.482	.333	.259	.395	
	090	2.236	1.888	2.588	.350	.277	.410	
	100	2.314	1.964	2.648	.364	.293	.423	
	110	2.664	2.296	3.018	.426	.361	.480	
	120	2.880	2.588	3.351	.474	.415	.525	
	130	3.280	2.885	3.678	.516	.460	.565	
	140	3.576	3.167	3.985	.553	.501	.600	
	150	3.873	3.458	4.305	.588	.538	.634	

表 11-57 观测与期望频数

单元计数和残差								
数字	agent	dose	主体数	观测的响应	期望的响应	残差	概率	
PROBIT	1	.410	50	6	6.769	-.769	.135	
	2	.580	48	16	16.170	-.170	.337	
	3	.710	46	24	24.852	-.852	.540	
	4	.890	49	42	38.916	3.084	.794	
	5	1.010	50	44	45.176	-1.176	.904	
	6	1.000	48	18	15.035	2.965	.313	
	7	1.310	48	34	37.198	-3.198	.775	
	8	1.480	49	47	45.301	1.699	.925	
	9	1.610	50	47	48.741	-1.741	.975	
	10	1.700	48	48	47.508	.492	.990	
	11	1.009	50	44	45.153	-1.153	.903	
	12	.886	49	42	38.762	3.238	.791	
	13	.708	46	24	24.712	-.712	.537	
	14	.580	48	16	16.214	-.214	.338	
	15	.415	50	6	7.019	-1.019	.140	

表 11-58 各组中位数效应比较值

		相对中位数强度估计值			对数转换的 95% 置信限度 <sup>a</sup>		
		95% 置信限度			估计		
(I) agent	(J) agent	估计	下限	上限	估计	下限	上限
PROBIT 1	2	.366	.246	.500	-.437	-.609	-.301
	3	1.001	.864	1.161	.001	-.063	.065
2	1	2.733	1.998	4.066	.437	.301	.609
	3	2.737	2.000	4.074	.437	.301	.610
3	2	.365	.245	.500	-.437	-.610	-.301
	1	.999	.861	1.157	-.001	-.065	.063

a. 对数底数 = 10

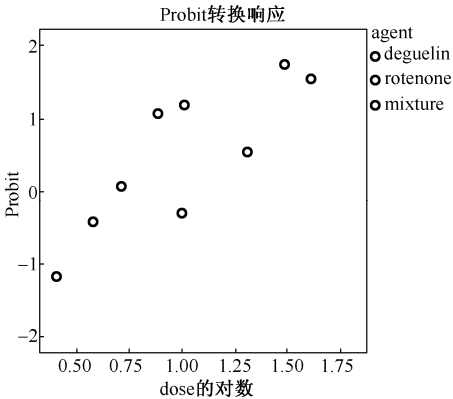


图 11-63 散点图

## 11.7 非线性回归

### 11.7.1 非线性模型

#### 1. 本质线性模型与本质非线性模型

$$y = e^{b_0 + b_1 x_1 + b_2 x_2 + e}$$

上式所表达的模型两边取自然对数，就可以写为

$$\text{Ln} y = b_0 + b_1 x_1 + b_2 x_2 + e$$

这种看起来非线性，但可以转换为线性的模型，称为本质线性模型。

当把一个模型转换为线性模型后，必须确保转换后的误差项也要满足所需的假设条件。例如，对于原始方程  $y = e^{bx} + e$ ，由于取对数后失去误差项  $e$ ，为了保证在转换后的模型中也存在误差项，原始方程式应写为

$$y = e^{bx+e} = e^{bx} e^e$$
$$y = b_0 + e^{b_1 x_1} + e^{b_2 x_2} + e^{b_3 x_3} + e$$

不能转换为线性的模型，称为本质非线性模型。在非线性回归过程中，必须首先估算将会应用到非线性模型中的起始值和参数值的范围，目的只是要将残差平方和减少到最小。本节解决本质非线性问题。

#### 2. 常用非线性模型

表 11-59 所示是已经得到公认并经常使用的非线性模型。

注意：不能随意套用。

表 11-59 常用非线性模型

名 称	模型表达式
Asymptotic	$b_1 + b_2 \exp(-b_3 x)$
Asymptotic	$b_1 - (b_2 b_3^x)$
Density	$(b_1 + b_2 x)^{(-1/b_3)}$
Gauss	$b_1 (1 - b_3 \exp(-b_2 x^2))$
Gompertz	$b_1 \exp[-b_2 \exp(-b_3 x)]$
Johnson-Schumacher	$b_1 \exp[-b_2 / (x + b_3)]$
Log-Modified	$(b_1 + b_3 x)^{b_2}$
Log-Logistic	$b_1 - \ln[1 + b_2 \exp(-b_3 x)]$
Metcherlich Law of Diminishing Returns	$b_1 + b_2 \exp(-b_3 x)$
Michaelis Menten	$b_1 x / (x + b_2)$
Morgan-Mercer-Florin	$(b_1 b_2 + b_3 x^{b_4}) / (b_2 + x^{b_4})$
Peal-Reed	$b_1 / \{1 + b_2 \exp[-(b_3 x + b_4 x^2 + b_5 x^3)]\}$
Ratio of Cubics	$(b_1 + b_2 x + b_3 x^2 + b_4 x^3) / (b_5 x^3)$
Ratio of Quadratics	$(b_1 + b_2 x + b_3 x^2) / (b_4 x^2)$
Richards	$b_1 / \{[1 + b_3 \exp(-b_2 x)]^{1/b_4}\}$
Verhulst	$b_1 / [1 + b_3 \exp(-b_2 x)]$
Von Bertalanffy	$[b_1^{(1-b_4)} - b_2 \exp(-b_3 x)]^{1/(1-b_4)}$
Weibull	$b_1 - b_2 \exp(-b_3 x^{b_4})$
Yield Density	$(b_1 + b_2 x + b_3 x^2)^{-1}$

3. 条件逻辑表达式

条件逻辑表达式应用于方程中或损失函数中。为了表达一个模型中或损失函数中的条件逻辑式，必须将几个不同条件的分段模型组合在一起。每个分段模型由逻辑表达式乘以逻辑表达式为真时的结果。例如，分段模型表示为

$$\hat{f}(x)=\begin{cases}0 & x\geqslant 0 \\ x & 0<x<1 \\ 1 & x\leqslant 1\end{cases}$$

这几个分段模型组合后的逻辑表达式为  $(x\leqslant 0)\cdot 0+(x>0\ \&\ x<1)\cdot x+(x\geqslant 1)\cdot 1$ ，因为逻辑表达式的值只能是 1(真)或 0(假)，因此：

如果  $x\leqslant 0$ ，以上结果为  $1\cdot 0+0\cdot x+0\cdot 1=0$ ；

如果  $0<x<1$ ，以上结果为  $0\cdot 0+1\cdot x+0\cdot 1=x$ ；

如果  $x\geqslant 1$ ，以上结果为  $0\cdot 0+0\cdot x+1\cdot 1=1$ 。

两个不等式之间必须由逻辑运算符连接。例如， $0<x<1$  必须写成  $(x>0\ \&\ x<1)$ 。

字符串表达式可用于逻辑表达式中。 $(sex='M')\cdot worth+(sex='F')\cdot 0.59\cdot worth$  的结果为：当变量 sex 值为 M 时变量 worth 的值，与变量 sex 值为 F 时变量 worth 的值乘 0.59%之和。

4. 损失函数

在非线性回归中，损失函数是对某统计量的运算法则，非线性回归过程以将其值最小化为原则进行非线性拟合。SPSS 默认根据最小残差平方和找出非线性模型。也可以自定义损失函数。

5. 参数约束

在数多的非线性模型中，参数必须限制在有意义的区间中。所谓约束是指在利用迭代方法求解的过程中对参数值的限制。可以首先使用线性约束，防止结果溢出。

- ① 线性约束：将参数乘以常数，该常数不能是其他参数或者自身。
- ② 非线性约束：其中至少一个参数与其他参数相乘或相除或者进行幂运算。

6. 数据要求

因变量和自变量应该是数值型变量。名义变量应该被重新编码为二分(哑)变量或者是其他类型的对比变量。同时要求定义的函数要尽可能精确地反映因变量与自变量之间的关系。

7. 估算初始值

即使模型是非常精确的，准确地确定参数的初始值也是非常重要的。为参数设置合适的初始值以保证正常、迅速收敛，同时避免解决方案范围小于实际范围。

- (1) 使用图形辅助确定参数取值范围，在研究的实际范围内确定初始值。
- (2) 根据确定的非线性方程的数学特性进行变换，结合图形辅助判断初始值范围。
- (3) 直接使用数值来替代某些参数，确定其他参数的取值范围，从而确定初始值。
- (4) 将数据转换后，使用线性关系模型确定初始值。通常联合使用上述几种方法。如果参数没有初始值，也不要仅仅将它们设置为 0，最好是将它们设置为预计要改变的值的的大小。如果忽略误差项，或许可以获得一个线性模型，并根据线性模型估算初始值。例如，模型  $y=e^{a+bx}+\varepsilon$ ，

如果忽视误差项 $\varepsilon$ , 并且在两边取对数, 获得模型  $\ln y = a + bx$ , 就可以利用线性模型来估计参数  $a$ 、 $b$  的值了。

(5) 利用非线性模型的属性估算初始值。有时能确定因变量在一定范围内的值。例如, 如果在模型  $y = e^{a+bx}$  中, 当  $x=0$  时,  $y=2$ , 就可以取  $\ln 2$  作为参数  $a$  的初始值。考虑当模型的值为最大值和最小值, 或当所有的自变量接近 0, 或其值接近无限大时的情况, 会对确定参数的起始值有帮助。

(6) 利用与参数同等数量的方程式, 可以解决参数的初始值问题。再看前面的例子, 可以解联立方程

$$\begin{cases} \ln y_1 = a + bx_1 \\ \ln y_2 = a + bx_2 \end{cases}$$

利用减法得  $\ln y_1 - \ln y_2 = bx_1 - bx_2$ , 解此方程式, 得参数  $b = \frac{\ln y_1 - \ln y_2}{x_1 - x_2}$ ,  $a = \ln y_1 - bx_1$ 。

## 11.7.2 非线性回归过程

(1) 按【分析→回归→非线性】顺序打开如图 11-64 所示的【非线性回归】主对话框。从源变量框中选择一个数值型变量作为因变量送入【因变量】框中。

(2) 在【模型表达式】框中输入合适的模型表达式, 其中应至少包括一个自变量。

① 将变量选入【模型表达式】框中。字符型变量仅能在逻辑表达式中使用。

② 定义模型表达式。从【函数组】框中选择需要的非线性函数送入【模型表达式】框中。从计算模板上选择数字或操作符, 组成模型表达式。

**注意:** 参数名不能与所选择的变量同名。

③ 定义模型参数。单击【参数】按钮, 打开如图 11-65 所示的【非线性回归: 参数】对话框。在【名称】框中输入参数名。在【初始值】框中输入尽量准确的初始值, 即尽可能接近期望值。定义一个, 单击【添加】按钮确定一个, 直到把所有参数定义完。选择某个参数, 单击【删除】按钮可将其剔除; 修改后单击【更改】按钮确认。定义或修改完成, 单击【继续】按钮返回主对话框。这里设置的参数以及初始值将在以后的分析中一直起作用。



图 11-64 【非线性回归】对话框



图 11-65 【非线性回归: 参数】对话框

如果前次运行非线性函数, 参数显示在【参数】框中。要使用这些参数作初始值, 在对话框中选择【使用上一分析的起始值】。如果修改了模型表达式, 则不能选择此项。

(3) 如果需要对【参数】框中的参数取值范围进行约束,单击【约束】按钮,打开如图 11-66 所示的对话框。

- ①【未约束】。默认对参数的值不限制。
- ②【定义参数约束】。定义对参数的限制。在【参数】框中选择需要约束的参数送入【定义参数约束】框。在逻辑运算符下拉列表中选择 $\leq$ 、 $\geq$ 、 $=$ 这 3 个中的任意一个,在最右上侧的框中输入常数。构成约束表达式后,单击【添加】按钮送入右下角的框中。选择表达式,单击【删除】按钮可将其删除;修改后单击【更改】按钮确认,显示新表达式。选择【继续】按钮返回主对话框。



图 11-66 【非线性回归: 参数约束】对话框

(4) 在非线性回归中,默认的损失函数是残差平方和。要自定义损失函数,在参数框中选择一个或多个参数,然后单击【损失】按钮,打开如图 11-67 所示的对话框。

- ①【残差平方和】。这是系统默认的损失函数。
- ②【用户定义的损失函数】。选择此项,输入自定义的损失函数。RESID\_表示残差; PRED\_表示预测值; 规定 RESID\_2 表示残差的平方和。

(5) 单击【选项】按钮,打开【非线性回归: 选项】对话框,见图 11-68,确定标准误的估计方法或者确定迭代过程停止的判定标准。



图 11-67 【非线性回归: 损失函数】对话框



图 11-68 【非线性回归: 选项】对话框

①【标准误的 Bootstrap 估计】。有放回地反复从原始数据集中提取相同容量的样本,来估算标准误。针对每一个样本建立相应的非线性回归模型,计算每个参数估计的标准误作为自举估计的标准差。原始数据的参数值作为每一个自举样本的初始值。

②【估计方法】栏。选择估计方法。

- 【序列二次编程(连续二次规划)】。适用于限制模型与非限制模型。如果确定了一个限制模型、定义了损失函数或选了自举估计,则自动选中该项。它利用双重迭代算法求解,每一步迭代建立一个二次规划算法,确定寻找的方向,并在选择的方向中发现一个新点,而损失函数对新点进行求值,直到寻找过程发生收敛。判定标准和精度选项如下:

A. 【最大迭代】框。输入最大迭代步数作为迭代停止的判定标准。

B. 【步长限制】框。输入一个正值作为参数向量长度的最大允许变化量。

C. 【最优性容差】下拉列表。选择最优容限，即目标函数的精度，也即有效位数。如果容限为 0.1E-6，有效数字为 6 位。最优容限值必须大于函数精度。

D. 【函数精度】下拉列表。选择小于最优容限并在 0~1 之间的数字作为目标函数精度。函数值较大时，作为相对精度；函数值较小时，作为绝对精度。

E. 【无限步长】下拉列表。在一步迭代过程中参数的变化大于设置值，迭代停止。

● 【Levenberg-Marquardt】。非线性约束模型的默认运算法则，如果确定了一个线性约束模型，或者定义了一个损失函数，或者选中标准误的自举估计，那么该选项不起作用。控制迭代停止的判定标准有：

A. 【最大迭代】。输入 Levenberg-Marquardt 算法中最大的迭代步数。

B. 【平方和收敛性】框。残差平方和的变化量小于设置值，迭代停止。

C. 【参数收敛性】。任何一个参数值的变化小于设置值，迭代停止。

后两项的默认值均为 1E-8。

(6) 单击【保存】按钮，打开如图 11-69 所示的【非线性回归：保存新变量】对话框。指定要生成的新变量。

① 【预测值】。因变量预测值，变量名为 Pred\_。

② 【残差】。变量名为 Resid。

③ 【导数】。为每一个模型参数保存导数，变量名为参数名前加前缀“d”。

④ 【损失函数值】。定义了损失函数，才会保存损失函数变量值，其变量名为 Loss\_。

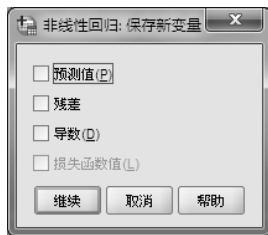


图 11-69 【非线性回归：保存新变量】对话框

### 11.7.3 非线性回归分析实例

【例 10】数据文件 data11-08 是美国 1790—1960 年人口变化的数据，人口单位为百万。

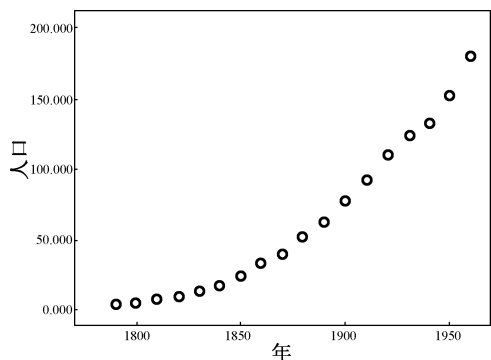


图 11-70 美国 1790—1960 年人口散点图

图 11-70 所示为人口与年份的散点图。根据经验，人口数量模型的建立经常使用 Logistic 模型，其方程为

$$y_i = \frac{c}{1 + e^{a+bt_i}} + e_i$$

式中， $y_i$  是在时间  $t_i$  时的人口数量； $e_i$  为误差项； $a$ 、 $b$  为参数。虽然通常模型对观测数据的拟合程度相当好，但有关误差项的独立性假设和常数项方差的假设却有可能被破坏。这是由于时间序列数据的误差项通常并不独立，误差项的大小有可能依数据总体的大小而变化。由于人口成长的模型不能被转换为线性模型，

因此选择非线性模型来估算模型的参数。

#### 1) 初始值的确定

本例利用简单的假设来确定初始值。在 Logistic 人口增长模型中，参数  $c$  为渐近线。任意选择距最大观测值不远的渐近线。本例最大人口值为 178，故选择 200 作  $c$  的初始值，然后依

据时间为 0 的人口值来估算参数  $a$  的值:

$$3.895 = \frac{200}{1 + e^{a+b \cdot 0}}$$

$$a = \ln\left(\frac{200}{3.895} - 1\right) = 3.9$$

接下来利用时间为 1 时的人口值来估算参数  $b$  的初始值:

$$5.267 = \frac{200}{1 + e^{b+3.9}}$$

$$b = \ln\left(\frac{200}{5.27} - 1\right) - 3.9 = -0.29$$

最终获得参数  $a$ 、 $b$  的初始值分别为 3.9、-0.29。

如果在确定初始值时没有非常明确的范围，可以先根据对函数的了解设定参数的初始值，再在【非线性回归：参数约束】对话框中设定参数的数值范围。这样可能达到最优回归的步数多一些，运行时间长一些。只要非参数模型选择正确，最终总能得到比较满意的结果。

- 2) 调用过程
- (1) 读取数据文件 data11-08，按【分析→回归→非线性】顺序打开主对话框。

(2) 将变量 pop 设置为因变量，送入【因变量】框中。

(3) 在【模型表达式】框中输入估计的模型表达式  $c/[1+2.718(a+b \cdot \text{decade})]$ 。

(4) 在【非线性回归：参数】框中根据前面计算结果，设定  $a \approx 3.9$ 、 $b \approx -0.29$ 、 $c = 200$ 。

(5) 在【非线性回归：保存新变量】对话框中选择【预测值】、【残差】选项。
- 3) 输出结果(见表 11-60~表 11-63，图 11-71、图 11-72)

表 11-60 每步迭代的残差平方和、参数值

迭代数 <sup>a</sup>	残差平方和	参数		
		a	b	c
1.0	2199.753	3.900	-.290	200.000
1.1	203.656	3.883	-.278	241.492
2.0	203.656	3.883	-.278	241.492
2.1	186.497	3.890	-.279	243.967
3.0	186.497	3.890	-.279	243.967
3.1	186.497	3.889	-.279	243.988
4.0	186.497	3.889	-.279	243.988
4.1	186.497	3.889	-.279	243.987

表 11-61 非线性模型统计量摘要

源	平方和	df	均方
回归	123053.531	3	41017.844
残差	186.497	15	12.433
未更正的总计	123240.028	18	
已更正的总计	53293.925	17	

因变量: 人口

a. R 方 = 1 - (残差平方和) / (已更正的平方和) = .997。

导数是通过数字计算的。

a. 主迭代数在小数左侧显示，次迭代数在小数右侧显示。

b. 由于连续残差平方和之间的相对减少量最多为 SSCON = 1.000E-008，因此在 8 模型评估和 4 导数评估之后，系统停止运行。

表 11-60 所示是迭代各步的残差平方和与参数  $a$ 、 $b$ 、 $c$  的估计值。每步迭代后，计算估算值的变化量。表的最后一部分表示在估算完 8 个模型、4 个导数后，由于两次迭代的最小残差平方和的减少量小于默认的收敛判定标准 1.E-08 而终止。

根据前面的计算，得出最终的回归方程为

$$y_i = \frac{243.99}{1 + e^{3.89 - 0.28t_i}}$$

表 11-61 所示为非线性模型统计量的摘要。它包括回归平方和、残差平方和，总平方和(因变量各观测值的平方和)(注：表中译成“未更正的总计”)、校正总平方和(因变量各观测值对均值的偏差平方和)。



$$R^2 = 1 - \frac{\text{Residual Sum of Squares}}{\text{Corrected Sum of Squares}} = 1 - \frac{186.497}{53293.925} = 1 - 0.0034 = 0.9966$$

表明模型对数据的拟合程度非常好。如果模型的拟合程度非常差， $R^2$  也可能为负值。  
图 11-71 是使用双轴 (Dual Axes) 图形功能完成的。第二个纵轴是预测值。从预测值与观测值的散点也可以看出拟合得很好。

**注意：**不能使用线性模型的检验方法检测非线性模型。即使模型非常正确，残差均值平方也不再是误差方差的无偏估计。为了应用的目的，仍可以比较残差方差和估算总方差，但是  $F$  统计量不能再用来对假设进行检验。

在非线性模型中不大可能获得每个参数精确的置信区间，大样本一般依靠渐近线的近似值进行估算。表 11-62 所示为各种参数估计值，表 11-63 所示为估计参数的渐近相关矩阵。

表 11-62 参数估计值

参数估计值				
参数	估计	标准误	95% 置信区间	
			下限	上限
a	3.889	.094	3.690	4.089
b	-.279	.016	-.312	-.246
c	243.987	17.968	205.690	282.285

表 11-63 估计参数的渐近相关矩阵

参数估计值的相关性			
	a	b	c
a	1.000	-.724	-.376
b	-.724	1.000	.904
c	-.376	.904	1.000

图 11-72 所示为残差对观测年代的散点图。观察图形可见，残差的方差随着时间的增长而增长。为计算预测值的渐近标准误和其他统计量，可以进行以残差为因变量的线性回归分析。

如果在表 11-63 中出现非常大的正值或者负值，很可能是由于模型中参数过多 (较少参数的模型就能很好地拟合数据)，相对来说观测的数量不足，但不说明模型不适合。

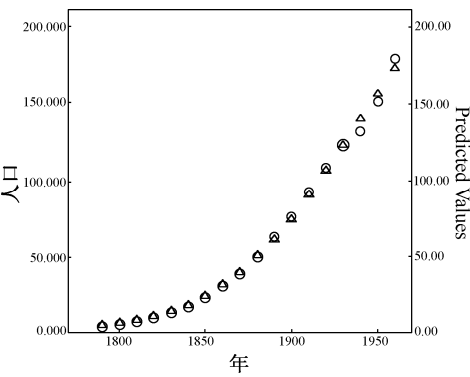


图 11-71 观测与预测值的散点图

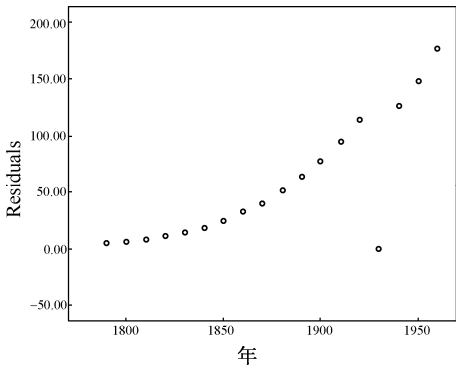


图 11-72 残差-年度散点图

## 11.8 加 权 回 归

### 11.8.1 加权回归的概念

回归模型为

$$y_i = b_0 + b_1x_{i1} + \cdots + b_px_{ip} + e_i$$

在前面介绍过的线性回归中，是用普通的最小二乘法 (OLS) 来建模的，它要求的前提条件是式中的误差项  $e_i$  服从均值为 0、方差为  $\delta^2$  的正态分布，也就是所有的观测在计算过程中具有相同的贡献。但对于某些观测的一些特性变异较其他观测大时，使用 OLS 就不能获得较好的模型。

如果上面模型中的误差项  $e_i$  服从均值为 0、方差为  $\sigma^2 x_i^w$  的正态分布，即因变量的方差与预测变量的值有关，换言之，如果它们的变异性是可以通过其他变量进行预测，就可以使用加权最小二乘法(WLS)来拟合线性回归模型。加权回归给出加权转换的范围，并得出最佳的权数值。

例如，考虑到由于高市值的股票较低市值的股票具有较高的变异性(价格的上下波动)，仅使用一般线性回归过程的 OLS 进行估算就不能很好地反映通货膨胀与失业率对变异性较大股票价格的影响，而 WLS 可以较好地解决这个问题。再如，健康研究中，各种治疗方法对病人住院时间长短的影响，很明显需要住院时间越长的病情，其表现的变异性就要比住院短的病人的病情所表现的变异性要大；产品研究中，工人的训练水平与产品质量之间的关系，产品质量越差，其变异性越大；社会学与犯罪学研究中，犯罪率较高的地区要比犯罪率较低的地区表现出更高的变异性。

1. 诊断与权重估计

(1) 图形。

图 11-73(a)所示的例子(数据文件 data11-09 来自 1981 年 DRVPER 和 SMITH)，图中只有两个变量  $x$  和  $y$ 。可以观察到因变量的变异性或分布随着自变量的增加而增加，这暗示着方差相同的假设已经遭到破坏，且最小平方方法不再是最佳解决方案了。

观察图 11-73(b)所示的预测值与残差散点图，可以得出相同的结论。

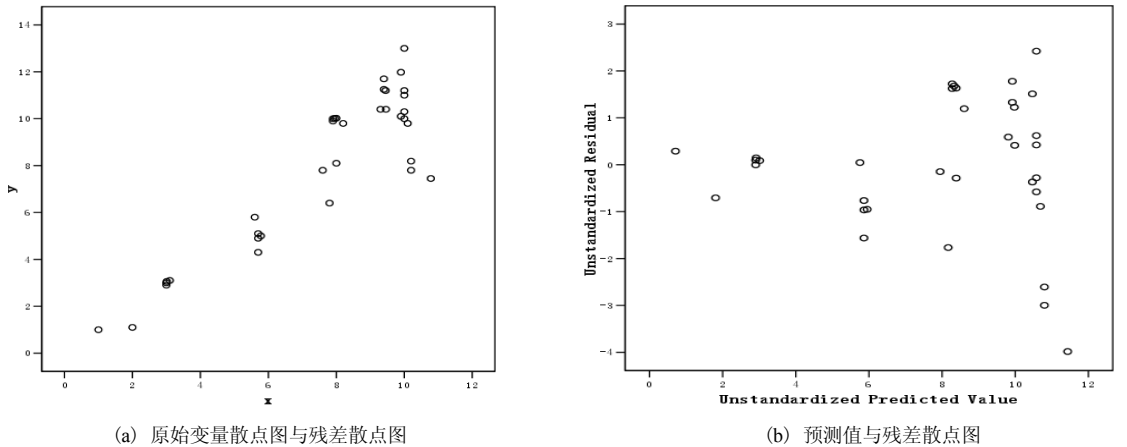


图 11-73 原始变量散点图与残差散点图

(2) 估计权重的方法。

① 由数据的复制集估计权重。为了使用加权最小平方方法来估计回归模型，将具有相同特点或近似特点的数据进行编组(数据的复制集)。这样就可以计算因变量相对于每一组具有不同特点的自变量的方差了。此时得到的方差的倒数就是权重。

② 由变量估计权重。如果认为因变量的方差与自变量或者其他变量之间存在关系，就可以使用 WLS 来估计权重。例如，研究收入与受教育程度之间的关系可知那些有研究生学历人员的工资变异要比那些没有获得学位的人员工资的变异高得多。

方差、变量、指数之间的关系如下： $\text{方差} \propto \text{变量}^{\text{指数}}$ 。可以指定指数值的范围或者一个增量，SPSS 将会规定的范围内估计所有指数值的对数似然比值，然后选择出具有最大似然比值的指数值。

2. 数据要求

自变量和因变量应该是数值型变量，类似宗教、民族和地区这样的分类变量应该被重新编码作为二分(哑)变量或其他类型的对比变量。加权变量必须是与因变量有关的数值型变量。对于自变量的每个值，要求因变量的分布必须是正态的。因变量和每一个自变量的关系应该是线性的，并且所有的观测应该是相互独立的。自变量取不同值时，因变量的方差不同，但是这些差异一定是可以根据加权变量预测出来的。

11.8.2 加权回归过程

- (1) 按【分析→回归→权重估计】顺序打开【权重估计】主对话框，见图 11-74。
- (2) 从左侧的源变量框中选择一个变量作为因变量进入【因变量】框中。
- (3) 从源变量框中选择一个或多个的自变量进入【自变量】框中作为自变量。
- (4) 从源变量框中选择一变量，将其选入【权重变量】框中，作为加权变量。观测数据的权重为  $1/wv^{power}$ ， $wv$  为加权变量， $power$  为加权变量指数。
- (5) 在【幂的范围】框中输入将在计算权重的过程中所使用的指数值的范围。第一个框中设定初始值，【到】框中设定结束值，【按】框中设定步长，应该保证(初始值-结束值)/步长≤150。指数值的范围必须在-6.5~7.5 之间。
- (6) 【在等式中包含常量】选项。模型中包括常数项。
- (7) 单击【选项】按钮，打开如图 11-75 所示的【权重估计：选项】对话框，确定在数据文件中保存的新变量，并确定方差和估测值的列表形式。

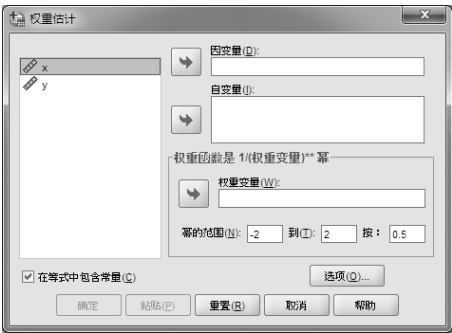


图 11-74 【权重估计】主对话框

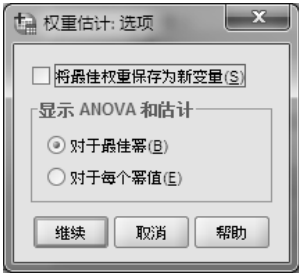


图 11-75 【权重估计：选项】对话框

- ① 【将最佳权重保存为新变量】。保存新变量。新变量的值是根据最大对数似然比函数计算的最佳权重值。变量名为 WGT\_n，n 是运行、生成这个新变量的序号。
- ② 【显示 ANOVA 和估计】栏，确定方差和估计值的输出形式。
  - 【对于最佳幂】。只输出最终的方差和最佳指数估计值。
  - 【对于每个幂值】。输出方差和所设置范围的指数值。
- (8) 单击【确定】按钮提交运算。

11.8.3 加权回归分析实例

【例 11】 数据文件 data11-09 有两个变量  $x$ 、 $y$ 。求以  $x$  为自变量、 $y$  为因本量的回归方程。

1) 操作步骤

- (1) 按【分析→回归→权重估计】顺序打开【权重估计】对话框。

(2) 选择变量  $y$  为因变量送入【因变量】框,  $x$  为自变量送入【自变量】框。 $x$  变量作为加权变量进入【权重变量】框。

(3) 设置加权的指数值, 初始值为 0, 结束值为 2.5, 步长为 0.1。

(4) 在【权重估计: 选项】对话框中选择【将最佳权重保存为新变量】选项, 保存每一个观测的权重值, 其他选项为默认设置。

(5) 单击【确定】按钮, 提交运算。结果输出见表 11-64~表 11-68。

表 11-64 说明自变量为  $x$ , 因变量为  $y$ , 按照 0.1 为步长的权值计算出的对数似然比结果如表所示。 $-55.543526$  为最大值, 因此得到的最佳指数值为 1.900。

表 11-65 所示为回归效果的统计量, 其中源变量为  $x$ , 因变量为  $y$ , 权重值为 1.9。

表 11-66 中的统计量说明模型对数据的拟合程度较好。还要看方差分析结果。

表 11-67 所示为对回归方程的方差分析。 $F$  值为 567.3, 显著水平值小于 0.01, 说明由回归解释的变异远远大于残差可解释的变异, 回归效果是比较好的。

表 11-68 所示是对回归方程中自变量  $x$  的系数为 0 的假设检验。T 检验的  $\text{Sig.} < 0.05$ , 拒绝  $x$  系数为 0 的假设, 也说明回归效果较好。所得方程式的最后结果为  $y = -0.283 + 1.077x$ 。

在数据窗中生成新变量 WGT\_1。

表 11-64 权值

对数似然值 <sup>b</sup>	
幂	
.000	-61.796
.100	-61.281
.200	-60.775
.300	-60.279
.400	-59.796
.500	-59.327
.600	-58.873
.700	-58.437
.800	-58.020
.900	-57.626
1.000	-57.255
1.100	-56.912
1.200	-56.598
1.300	-56.319
1.400	-56.075
1.500	-55.872
1.600	-55.714
1.700	-55.603
1.800	-55.545
1.900	-55.544 <sup>a</sup>
2.000	-55.603
2.100	-55.726
2.200	-55.918
2.300	-56.182
2.400	-56.521
2.500	-56.938

a. 选择对应幂以用于进一步分析, 因为它可以使对数似然函数最大化。

b. 因变量:  $y$ , 源变量:  $x$

表 11-65 模型描述

模型描述	
因变量	$y$
自变量	$x$
权重	源
幂值	1.900
模型: MOD_2.	

表 11-66 模型综述

模型摘要	
复相关系数	.972
R 方	.945
调整 R 方	.943
估计的标准误	.197
对数似然函数值	-55.544

表 11-67 方差分析

ANOVA					
	平方和	df	均方	F	Sig.
回归	22.116	1	22.116	567.319	.000
残差	1.286	33	.039		
总计	23.403	34			

表 11-68 模型参数及各种统计量

	未标准化系数		标准化系数		t	Sig.
	B	标准误	试用版	标准误		
(常数)	-.283	.194			-1.457	.155
$x$	1.077	.045	.972	.041	23.818	.000

2) 与一般线性回归方法进行比较

(1) 以同一数据文件进行一般线性加权回归分析和一般线性不加权回归分析。

① 一般线性回归: 选择  $x$  作为自变量,  $y$  作为因变量进行一次回归。

② 一般线性加权回归：选择  $x$  作为自变量， $y$  作为因变量，新变量 WGT\_1 作为加权变量送入【权重变量】框中。不选【在等式中包含常量】选项。

对比表 11-69 和表 11-70，一般线性回归的  $R$  值为 0.905， $R^2$  为 0.819，明显小于加权回归的对应值。这说明方程式加权后的效果是十分明显的。

表 11-69 一般线性回归小结

模型汇总				
模型	R	R 方	调整 R 方	标准 估计的误差
1	.905 <sup>a</sup>	.819	.814	1.45535

a. 预测变量: (常量),  $x$ 。

表 11-70 加权线性回归小结

模型摘要	
复相关系数	.986
R 方 <sup>a</sup>	.973
调整 R 方	.972
估计的标准误	1.440
对数似然函数值	-61.925

a. 对于通过原点的回归（无截距模型），R 方将测量因变量中有回归所解释原点的变异性的比例。对于包括截距的模型，这个指标不可与 R 方进行比较。

(2) 两种回归方法的方差分析如表 11-71、表 11-72 所示。系数检验结果如表 11-73、表 11-74 所示。

比较表 11-73、表 11-74，发现 WLS 和 OLS 的斜率(1.051、1.096)没有大的差别，但它们的标准误变化较大：斜率  $b$  的标准误分别为 0.090、0.030。

表 11-71 一般线性回归方差分析

Anova <sup>a</sup>						
模型		平方和	df	均方	F	Sig.
1	回归	316.433	1	316.433	149.399	.000 <sup>b</sup>
	残差	69.895	33	2.118		
	总计	386.329	34			

a. 因变量:  $y$

b. 预测变量: (常量),  $x$ 。

表 11-72 加权线性回归的方差分析

ANOVA					
	平方和	df	均方	F	Sig.
回归	2509.034	1	2509.034	1209.995	.000
残差	70.502	34	2.074		
总计	2579.536 <sup>a</sup>	35			

a. 此总平方和对于常量不正确，因为在通过原点的回归中常量为零。

表 11-73 一般线性回归的系数检验

系数 <sup>a</sup>						
模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	-.387	.722		-.535	.596
	x	1.096	.090	.905	12.223	.000

a. 因变量:  $y$

表 11-74 加权回归的系数检验

系数						
	未标准化系数		标准化系数		t	Sig.
	B	标准误	试用版	标准误		
x	1.051	.030	.986	.028	34.785	.000

为了进一步验证加权模型的效果，作转换后的预测值与残差的散点图。需要注意，在线性回归过程中首先保存预测值和残差到数据文件中，在绘制散点图之前对它们进行转换，转换的方法是它们本身乘以加权变量的 1/2 次方。

绘制的图形如图 11-76 所示，可以看出转换后的预测值对残差值的散点图的喇叭形状比图 11-73(b)有了改善。说明 WLS 获得了一定的效果。

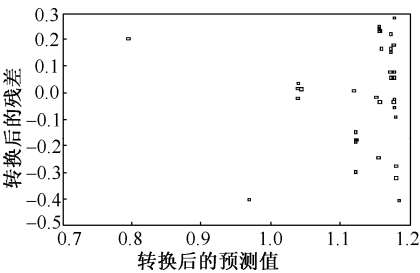


图 11-76 转换后的预测值对残差值的散点图

## 11.9 二阶段最小二乘法

### 11.9.1 二阶段最小二乘法的概念

#### 1. 概述

在前面介绍的一般线性回归模型中,总是假定自变量与模型中的误差项之间是不存在线性相关的,此时,用最小二乘法可获取回归系数无偏且一致的估计。但是在计量经济的模型中,这种假定并非总能得到满足。

例如,在用商品的价格和消费者的收入作为自变量建立对该商品的需求为因变量的回归模型时,由于价格和需求互相具有倒数作用关系,即价格可以影响需求,而需求也可以影响价格,回归模型中的误差项与商品的价格之间就会存在相关,在这种情况下用最小二乘法建立的线性模型已不再是最佳的模型估计,因为此时得到的回归系数,在理论上可证它是有偏且不一致的。

在计量经济模型的研究中,习惯上称线性模型中的自变量为解释变量,称因变量为被解释变量,而将模型中的误差项称为扰动项。当解释变量与模型中的扰动项之间存在线性相关时,称该解释变量为内生变量;而当解释变量与模型中的扰动项线性无关时,称它为外生变量。在上面的例子中,价格变量为内生变量,而收入变量就是外生变量。

由此可知,仅当在线性回归模型中所有自变量均为外生变量时,方可用最小二乘法来估计回归系数,从而得到最小二乘法下的最佳回归模型。

为解决线性模型中存在的内生性,即有内生变量时模型系数的估计,在 20 世纪 50 年代,H.泰尔提出了分两步来解决模型内生性问题的二阶段最小二乘法。

第一步,先寻找与模型中的误差项不存在线性相关的变量如在上例中的消费者的收入和商品的滞后价格,用它们作为当前商品价格的自变量构建线性模型。由于这些变量在模型估计过程中被作为工具使用,以替代模型中与误差项相关的内生变量(当前商品价格),故将其称为工具变量。作为工具变量必须满足四个条件:①与所替代的内生变量高度相关;②与随机误差项不相关;③与模型中其他自变量不相关;④同一模型中需要引入多个工具变量时,这些工具变量之间不相关。这样,用来估计当前商品价格的这两个自变量与其模型中的误差项,及其两个变量自身之间都不存在线性相关,故可用最小二乘估计来获取最佳线性模型,从而得到当前商品价格的预测值。

第二步,用计算得到的当前商品价格的预测值作为当前商品价格的替代值,以此为自变量,来作因变量商品的需求的线性回归分析。由于这个计算得到的值是来自于与模型误差不相关的变量,因此用它来预测商品需求时,与当前模型中的误差项也不存在线性相关,就可以再次使用最小二乘法来估计因变量的最佳线性模型。

需要注意的是,工具变量并没有替代模型中的解释变量,只是在估计过程中作为“工具”被使用。在 SPSS 的二阶段最小二乘法过程的界面中,工具变量还简称为工具,因此,在工具框中需要输入的是在二阶段最小二乘法的第一步中用来计算内生变量预测值的变量。同样的变量可以出现在解释变量和工具列表框中。工具变量的数量至少与解释变量数量一样多。如果解释变量列表和工具变量列表相同,则结果与线性回归程序的结果相同。解释变量没被指定为工具变量的被当作内生变量。通常,在解释变量列表中的所有外生变量也被指定为工具变量。

## 2. 二阶段最小二乘法对数据的要求

(1) 因变量和自变量必须是定量变量。如果是分类变量,则需先要重新编码为二分类的哑变量或其他类型的对比变量。内生解释变量也应是定量变量(非分类变量)。

(2) 对于自变量的每个值,因变量的分布必须为正态分布。对于自变量的所有值,因变量分布的方差相等。

(3) 因变量和每个自变量之间呈线性关系。

## 3. 二阶段最小二乘法回归模型

(1) 模型。二阶段最小二乘法回归模型为

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} = [\mathbf{Z}_1, \mathbf{Z}_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\varepsilon}$$

$$\mathbf{Z}_1 = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\delta}$$

式中,  $\mathbf{y} = [\mathbf{Z}_1, \mathbf{Z}_2]$ ;  $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$ ;  $\boldsymbol{\varepsilon}$  和  $\boldsymbol{\delta}$  是各自具有均值为 0、协方差矩阵为  $\sigma^2 \mathbf{I}_n$  及  $\xi^2 \mathbf{I}_n$  的扰动, 即随机误差。

其中,  $\mathbf{y}$  是由因变量的一个样本数据组成的  $n \times 1$  维向量;  $\mathbf{Z}$  是观测预测变量的  $n \times p$  维矩阵;  $\boldsymbol{\beta}$  是  $p \times 1$  维的参数向量;  $\mathbf{X}$  是元素  $X_{ij}$  的  $n \times 1$  维矩阵,  $X_{ij}$  表示  $i$  个样品的第  $j$  个工具变量的观测值;  $\mathbf{Z}_1$  是  $\mathbf{Z}$  的  $n \times p_1$  维子矩阵, 表示观测的内生变量;  $\mathbf{Z}_2$  是  $\mathbf{Z}$  的  $n \times p_2$  维子矩阵, 表示观测的外生变量;  $n$  为样品量;  $p$  为预测变量个数;  $p_1$  为预测变量中的内生变量数;  $p_2$  为预测变量中的外生变量数;  $\boldsymbol{\beta}_1$  为  $\boldsymbol{\beta}$  中与  $\mathbf{Z}_1$  有关的参数的子向量;  $\boldsymbol{\beta}_2$  为  $\boldsymbol{\beta}$  中与  $\mathbf{Z}_2$  有关的参数的子向量。

(2) 估计量的计算方法。SPSS 中用到的估计技术是 Theil 在 1953 年发现的。首先在模型的等式两边左乘  $\mathbf{X}'$ , 得到

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{Z}\boldsymbol{\beta} + \mathbf{X}'\boldsymbol{\varepsilon}$$

因为干扰向量具有均值 0 和协方差矩阵  $\sigma^2(\mathbf{X}\mathbf{X})$ , 所以  $(\mathbf{X}\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\boldsymbol{\varepsilon}$  将有协方差矩阵  $\sigma^2 \mathbf{I}_n$ 。因此, 在上述等式两边乘以  $(\mathbf{X}\mathbf{X})^{-\frac{1}{2}}$  会得到多元线性模型

$$(\mathbf{X}\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{y} = (\mathbf{X}\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\mathbf{Z}\boldsymbol{\beta} + (\mathbf{X}\mathbf{X})^{-\frac{1}{2}}\mathbf{X}'\boldsymbol{\varepsilon}$$

$\boldsymbol{\beta}$  的普通最小二乘法估计量  $\hat{\boldsymbol{\beta}}$  为

$$\hat{\boldsymbol{\beta}} = [\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

在用普通最小二乘法进行的回归建模中, 所要用到的其他统计量方面的计算方法请参见前面相关各章中的内容。

### 11.9.2 二阶段最小二乘法过程

(1) 按【分析→回归→二阶段最小二乘法】顺序打开【二阶段最小二乘法】主对话框, 见图 11-77。

(2) 从左侧的源变量框中选择一个变量作为因变量进入【因变量】框。

(3) 从源变量框中选择一个或多个用来对因变量建模的自变量进入【解释变量】框。

(4) 从源变量框中选择一个或多个用来预测内生变量的工具变量, 将其选入【工具变量】框。

(5) 【在等式中包含常量】选项。模型中包括常数项。

(6) 单击【选项】按钮，打开如图 11-78 所示的【二阶段最小二乘法：选项】对话框，确定在数据文件中保存的新变量，并确定方差和估测值的列表形式。



图 11-77 【二阶段最小二乘法】主对话框

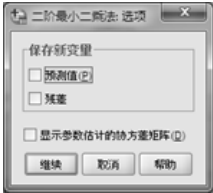


图 11-78 【二阶段最小二乘法：选项】对话框

- ①【保存新变量】框。可向当前文件中添加新变量。
    - 【预测值】。在活动文件中自动添加名为 FIT\_n 的新变量。n 是运行、生成这个新变量的序号。
    - 【残差】。在活动文件中自动添加名为 ERR\_n 的新变量。n 是运行、生成这个新变量的序号。
  - ②【显示参数估计的协方差矩阵】选项。在输出窗中输出参数估计的协方差矩阵。
- (7) 单击【确定】按钮提交运算。

11.9.3 二阶段最小二乘法分析实例

【例 12】 在数据文件 data11-10 中有工龄 workingage、年龄 age、当前工资 salary、初始工资 salbegin、受教育水平 educ、父亲受教育水平 Feduc、母亲受教育水平 Meduc、当前工资的以 10 为底的自然对数转换值 LGsalary，初始工资的以 10 为底的自然对数转换值 LGsalbegin 等变量，建立以 LGsalary 为因变量，以 LGsalbegin、workingage、age、educ 为自变量建立线性回归模型。

由于按照一般的常理，工资与学历(受教育水平)之间存在双向作用，受教育水平越高，一般年薪也会越高，而年薪达到一定数量后，就会有实力进一步攻读更高的学位，取得更高的学历，因此，线性回归模型中的受教育水平 educ 变量可当作内生变量来看待。由此可知，本例建立线性回归模型时，不能直接使用以最小二乘法作为算法基础的一般线性回归过程来建模，而需要用到二阶段最小二乘法过程来建模。

由于所要建立的模型以 LGsalbegin、workingage、age、educ 为自变量，以 LGsalary 为因变量，因此，在【因变量】框中需要输入的是 LGsalary。而在【解释变量】框中输入的是 LGsalbegin、workingage、age、educ。

由于受教育水平 educ，受到父亲受教育水平 Feduc、母亲受教育水平 Meduc 的影响，但反过来却不起作用，而且这两个变量与被调查对象的当前工资之间也无双向作用关系，因此可作为受教育水平 educ 变量的工具变量。另外，还将工龄 workingage、年龄 age、初始工资的以 10 为底的自然对数转换值 LGsalbegin 作为外生变量，一并用来作为预测受教育水平 educ 的自变量。

因此，在【工具变量】框中需要输入的变量为工龄 workingage、年龄 age、初始工资的以 10 为底的自然对数转换值 LGsalbegin、父亲受教育水平 Feduc、母亲受教育水平 Meduc。

通过对照进入【解释变量】框和【工具变量】框中的变量名列表可知，在【解释变量】框中受教育水平 educ 变量没有出现在【工具变量】框变量名列表中，因此，可以确定它是一个内生变量。这说明这样的设置正是本例所需的。



故在 SPSS 中的操作步骤如下:

- (1) 打开数据文件 data11-10, 按【分析→回归→两阶最小二乘法】顺序打开【二阶段最小二乘法】对话框。
- (2) 在左侧的源变量框中, 选择 LGsalary 为因变量送入因变量框, 选择 LGsalbegin、workingage、age、educ 为解释变量送入【解释变量】框, 选择 workingage、age、LGsalbegin、Feduc、Meduc 作为工具变量送入【工具变量】框中。
- (3) 在【二阶段最小二乘法: 选项】对话框中选择保存为新变量框下的预测值选项及显示参数估计的协方差矩阵选项。
- (4) 单击【确定】按钮, 提交运算。输出结果见表 11-75~表 11-80。

表 11-75 模型概述

模型描述		
		变量类型
方程 1	LGsalary	因变量
	workingage	预测值与工具
	age	预测值与工具
	educ	预测值
	LGsalbegin	预测值与工具
	Feduc	工具
	Meduc	工具
MOD_1		

表 11-76 模型统计量

模型汇总		
方程 1	复相关系数	.941
	R 方	.885
	调整 R 方	.880
	估计的标准误	.065

表 11-75 所示为参与建模的各变量的名称及在模型中的角色。表中变量类型为预测值的变量是所求模型中的内生变量(educ), 预测值与工具类型的变量是模型中的外生变量。同时, 工具及预测值与工具类型的变量是内生变量 educ 的预测变量(即自变量)。

表 11-76 给出了拟合优度统计量, 其中复相关系数为 0.941, 决定系数  $R^2$  为 0.885, 调整后的  $R^2$  为 0.880, 估计的标准误为 0.065。表明该回归方程可以解释总变异的 88.5%, 因此回归模型效果较好。

表 11-77 所示为对回归方程的方差分析结果。由于  $p = 0.000 < 0.05$ , 因此, 回归方程有统计学上的显著性意义, 可以用线性回归来描述 LGsalary 与相关因素的关系。

表 11-77 方差分析表

ANOVA					
	平方和	df	均方	F	Sig.
方程 1 回归	2.481	3	.827	194.612	.000
残差	.323	76	.004		
总计	2.804	79			

表 11-78 回归方程的系数及检验

	系数		Beta	t	Sig.
	未标准化系数	标准误			
方程 1 (常数)	.519	.217		2.391	.019
age	-.003	.001	-.150	-3.843	.000
educ	.008	.004	.117	2.257	.027
LGsalbegin	.960	.059	.851	16.391	.000

由表 11-78 可见, 回归系数  $p$  值均小于 0.05, 因此得到下列回归方程:  
$$LGsalary = 0.519 - 0.03age + 0.008educ + 0.960LGsalbegin$$

从表 11-79 可见, 工龄 workingage 变量没有进入方程, 其偏相关系数为 0.000, 最小容忍度为  $3.608 \times 10^{-16}$ , 远小于 0.1, 说明当该变量进入回归方程时共线性十分严重, 因此, 必须从回归方程中剔除。

表 11-79 排除出方程的变量

排除的变量					
	Beta In	偏相关	最小容忍	t	Sig.
方程 1 workingage	5.432	.000	3.608E-016	2.633E-006	1.000

表 11-80 所示是进入回归方程的 3 个变量之间的相关系数及协方差矩阵。  
另外，要求在文件中保存的因变量的预测值被存放在变量名 FIT\_1 的列中。

表 11-80 变量间的相关系数及协方差矩阵

系数相关性				
		age	educ	LGsalbegin
方程 1	相关性	age	1.000	.019
		educ	.019	1.000
		LGsalbegin	-.031	-.661
	协方差	age	5.420E-007	5.091E-008
		educ	5.091E-008	1.352E-005
		LGsalbegin	-1.342E-006	.000

11.10 最优尺度回归

11.10.1 最优尺度回归的概念

1. 概述

在标准线性回归分析中，回归方程中的变量通常是定量的，即使是名义变量，也会重新编码为二元变量或对比变量。而且对应于自变量的值，因变量须服从正态分布。建立在最小二乘法基础上的线性回归方程可以使得响应变量(因变量)和预测变量(自变量)的加权组合之间的平方差之和达到最小。分类变量可用来对观测进行分组。估计的系数反映了预测变量的变化对响应变量变化的影响程度。对于预测变量值的任何组合都可以预测响应变量的值。

但在调查研究中，尤其在问卷调查中，取得的多数资料不是定量变量，而是名义或有序的分类资料。例如，某服装制造商为了解不同性别、不同职业、不同学历的消费者 对服装颜色的偏好，在目标市场的消费者人群中开展有关这方面问题的问卷调查，希望建立起不同性别、不同职业、不同学历的消费者 对服装颜色的偏好的统计模型。在这个调查中，性别(男、女)、职业(学生、公务员、公司职员、自由职业者，其他)、颜色偏好(黑色、白色、红色、黄色、蓝色、其他)均为名义变量，学历(研究生、大学、高中、初中、其他)为有序变量。

由于这些分类变量的各个类别还没有做量化处理，因此，直接用标准的线性回归分析方法进行线性建模显然是不合适的。因而，首先想到的是对这些分类变量的各类别用数值编码的方式进行量化处理，如对性别变量，用 1 代表男，用 2 代表女；对职业变量用 1 代表学生、用 2 代表公务员、用 3 代表公司职员、用 4 代表自由职业者，用 5 代表其他；对学历变量用 1 表示研究生、用 2 表示大学、用 3 表示高中、用 4 表示初中、用 5 表示其他；对颜色偏好变量用 1 表示黑色、用 2 表示白色、用 3 表示红色、用 4 表示黄色、用 5 表示蓝色、用 6 表示其他。这样，一种可供选择的方法是用基于分类预测变量上的响应(因变量)来评价分类预测变量的回归，从而可为每个分类变量分别估计一个系数。但是，对于分类变量，类别值是任意的，如对颜色偏好变量用 1 表示黑色、用 3 表示白色、用 4 表示红色、用 8 表示黄色、用 9 表示蓝色、用 16 表示其他，也是可以的。但以不同方式编码类别将会产生不同的回归系数，这样在对同样几个变量的分析进行比较时，难度就增大了，有时甚至无法给出解释。

上述通过对分类变量用任意的数值进行编码方式的量化处理，并不能保证是最优的。因为对于名义分类变量，类别间的差异如何很难探索；对于有序分类变量，类别间的差异不一定相等；即使变量均为连续型变量，它们之间的联系也有可能为某种曲线，直接按线性结构来拟合

也不一定合适。因此,当需要用线性方程来表达某个分类变量与其他变量(包括分类变量)的关系时,需要用到一种新的回归方法,它就是使用交替最小二乘法的最优尺度分类回归法(CATREG),简称最优尺度回归。

要使用最优尺度回归分析法对分类变量进行回归分析时,首先必须对分类变量进行量化处理。分类变量的量化,并非简单地用正整数对其类别进行编码,而是要使用优化的尺度来量化转换分类变量为连续型数值变量,以使量化转换后的分类变量能用与数值型变量相同的方式进行处理。

所谓优化,是指在量化过程中,要使得对分类变量的量化结果能反映其初始类别的特征,并且保证变换后各变量间的关系为线性关系。这种优化是建立在分析分类变量的每个分类值对于因变量的影响程度的基础上,采用一定的非线性变换方法进行反复迭代,通过同时地尺度化名义变量、有序变量和数值变量,从而对原始变量的每个值都赋予一个最佳的量化数值。这样可用标准线性回归方法对转换后的变量进行回归分析,从而得到一个最佳的回归方程。

## 2. 最优尺度回归对数据的要求

(1) 最优尺度回归可对分类指示符变量进行运算。分类指示符应为正整数。最优尺度回归过程对自变量没有测度类型限制,对因变量也不作分布假定。

(2) 在最优尺度回归过程的离散化对话框中,可使用连续的正整数对名义、有序变量进行重新编码。使用 1 作为每个分类变量的起始点。如果变量已经是数值型变量,则不再重新编码。对有小数值的变量和有字符串的变量可通过重新编码的方法将其转换成正整数。

(3) 最优尺度回归只能设置 1 个因变量(多数情况下为分类变量),但最多可设置 200 个自变量。数据中至少要有 3 个有效观测,并且有效观测的数目必须超过自变量数加 1。

## 3. 分类变量离散化处理的方法

所谓分类变量的离散化处理是指对分类变量进行重新编码。它是在未加权的原始数据上进行的。其目的是使字符串变量转换成虚拟变量的编码方式,使得任意的多分类变量的类别值离散化为近似服从正态分布或均匀分布的类别值。常用的方法有:

(1) 乘法。首先,对原始变量进行标准化处理;然后,用标准化的值乘以 10,四舍五入取整,再加上一个值使得其最小值为 1。

(2) 赋秩法。用个案的秩次来代替原变量的值。

(3) 分组到服从正态分布的指定数量的类别中。首先,对原始变量进行标准化处理;然后,使用依据奥地利统计学家 Max (1960) 定义的间隔对观测值分类。

(4) 分组到服从均匀分布的指定数量的类别中。首先按除以指定的类别数,四舍五入取整,计算得到目标频数;然后原始类别被分到划分的组别中,如此以致组别中的频数尽可能接近目标频数。

(5) 分到与规定大小等间隔的组中。首先将间隔定义为最小值+间距,最小值+2 倍间距,等等;然后,具有第  $K$  个间隔值的样品被分到  $K$  类。

## 4. 最优尺度回归分析中用到的一些基本术语

在最优尺度回归分析中,由于大量使用线性规划中的术语和算法,因此出现了许多新的术语,为便于读者对本节内容的理解,特将与统计方法有关的主要的术语归纳如下,其他未列出的术语请读者参阅有关线性规划的书籍。

(1) 约束。受到一定条件的限制,不能超出限定的范围。

(2) 样条与样条约束。

① 样条是样条函数的简称。它是对若干个点进行曲线拟合的一种方式。如 3 次样条函数就是要求每两个点之间由 3 次曲线连接,而且在两段曲线的连接点(内点)上一阶、二阶导数相等,或者说在该点上曲线的切线和凹凸都相同。

② 样条约束。受到某个给定样条函数的限制。

③ 如果除受到给定的样条函数的制约外,还受到函数的单调性(递增或递减)限制,则称其受到样条及单调性约束。

④ 样条有序水平。一种用样条曲线拟合有序数据的方法。

(3) 目标函数与损失函数。目标函数是线性规划中对函数的另一种称谓。它是指所关心的目标(某一变量)与相关的因素(某些变量)的函数关系,也就是将目标表示为未知变量的线性表达式。与一般函数不同的是,在线性规划中,除对目标函数中出现的变量有一定的条件限制(即约束)外,还对目标函数有最大化或最小化的要求,以此来确定目标函数与未知变量的表达式。

损失函数是一种衡量损失和错误(这种损失与“错误地”估计有关,如费用或者设备的损失)程度的函数。

(4) 惩罚函数法。是线性规划中应用广泛且极为有效的间接解法,又称为序列无约束极小化方法(SUMT)。该方法通过将原约束化问题中的等式和不等式约束函数加权处理后与原目标函数结合,得到新的目标函数(称为惩罚函数)。这样可将原问题转换为新的无约束优化问题,求解该新的无约束优化问题,间接得到原约束化问题的最优解。

在目标函数的可行域中,是不需要增加任何限制的,一旦在计算中自变量越出可行域,就需要对它进行限制,即“惩罚”,因此,常把这样构造出来的函数称为惩罚函数。惩罚函数的作用是减少求解目标函数中的运算量,使得迭代计算中的收敛速度更快。

(5) 邻回归。邻回归分析是 1962 年由 Heer 首先提出的,1970 年后又与 Kennard 合作,系统地做了发展。它是一种改进的最小二乘法。

在一般线性回归中,用最小二乘法得到的回归参数  $\beta$  的估计为  $\hat{\beta} = (X'X)^{-1}X'Y$ 。其中,  $X$  是回归设计矩阵( $p$  个自变量  $X$  经标准化处理后的数据);  $X'$  为  $X$  的转置矩阵;  $X'X$  为相关系数矩阵;  $(X'X)^{-1}$  是相关系数矩阵的逆矩阵;  $Y$  为因变量向量。在线性规划中,回归参数  $\beta$  的

估计也可以用  $\hat{\beta}^{LS} = \arg \min \sum_{i=1}^n \left( Y_j - \sum_{j=1}^p X_{ij} \beta_j \right)^2$  来表示。其中,  $\arg \min$  表示使目标函数取最小值

时的变量值;  $\sum_{i=1}^n \left( Y_j - \sum_{j=1}^p X_{ij} \beta_j \right)^2$  表示均方误差。

当自变量间的相关性较大,存在共线性时,均方误差将变得很大,使用最小二乘法得到的系数估计已不再是无偏估计,因为此时相关系数矩阵的行列式的值近似于 0,存在奇异性,有可能使得相关矩阵的逆矩阵不存在,因此,此时需用改良的邻回归分析法。

邻回归的回归参数  $\beta$  的估计为  $\hat{\beta}^{\text{ridge}} = (X'X + \lambda I)^{-1}X'Y$ 。其中,  $\lambda$  是大于 0 的一个参数,  $I$  是  $p \times p$  的单位向量。由此可见,邻回归分析法实际上是在标准回归系数的计算公式中,通过在相关系数矩阵的每个对角线元素上加上一个常数  $\lambda$ ,来人为地增加了每个变量的变化范围,这可以改善最小二乘法估计中对回归系数估计的不稳定性。由于这个常数  $\lambda$  起到了限制相关矩阵的行列式的值近似于 0 的这种奇异性(也就是相关矩阵的逆矩阵不存在),因此,也称这样的  $\lambda$  值为惩罚值。研究表明,  $\lambda$  值不能取得太大。当  $\lambda$  值为 0 时,就还原为一般线性回归分析。

岭回归的回归参数  $\beta$  的估计还可表示为  $\hat{\beta}^{\text{ridge}} = \arg \min \left[ \sum_{i=1}^n \left( Y_j - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right]$ , 它

被称为岭回归的惩罚残差平方和, 而  $\lambda \sum_{j=1}^p \beta_j^2$  被称为惩罚项, 称其为 L2 惩罚。

(6) Lasso 回归。“Lasso”一词在 SPSS 中被汉化为“套索”, 是 Tibshirani (1996) 提出的一种关于线性回归的新方法。Lasso 是 The Least Absolute Shrinkage and Selectionator operator 的缩写。Lasso 回归的回归参数  $\beta$  的估计为  $\hat{\beta}^{\text{Lasso}} = \arg \min \left[ \sum_{i=1}^n \left\| Y_j - \sum_{j=1}^p X_{ij} \beta_j \right\|^2 + \lambda \sum_{j=1}^p |\beta_j| \right]$ , 它是在一

般线性最小二乘法的前提下加了约束, 使回归系数的绝对值之和小于某个常数, 从而通过构造一个惩罚函数获得一个精炼的模型; 使得该回归模型得出的一些变量的回归系数为零, 得到解释力较强的模型。从而实现了变量集精简的目的。它是一种处理具有复共线性数据的有偏估计。

在 Lasso 回归中用惩罚项  $\lambda \sum_{j=1}^p |\beta_j|$  来压缩模型系数, 因此也称为 L1 惩罚。

(7) 弹性网回归。在处理高维低样本的微阵列数据时, Zou 和 Hastie (2005) 针对一组具有复共线性的变量对因变量的影响中, 在 Lasso 的基础上引入了系数的二次惩罚, 提出了弹性网技术。这种方法不仅能有效地进行模型选择, 而且能处理自变量数目大于样本量的问题。

假定在数据集中有  $n$  次观测和  $p$  个变量。又假定因变量  $y$  是中心化的 (即预先用其原始观测值减去其均值作过变换处理), 预测变量  $X$  是标准化的 (即预先用其原始观测值减去其均值再除以其标准差作过变换处理)。

对于固定的两个非负数  $\lambda_1$ 、 $\lambda_2$ , 弹性网的目标函数为

$$L(\lambda_1, \lambda_2, \beta) = |y - x\beta|^2 + \lambda_1 |\beta|^2 + \lambda_2 |\beta|_1$$

式中,  $\beta$  为回归系数,  $|\beta|^2 = \sum_{j=1}^p \beta_j^2$ ,  $|\beta|_1 = \sum_{j=1}^p |\beta_j|$ ,  $j=1, 2, \dots, p$ 。  $\lambda_1 |\beta|^2 + \lambda_2 |\beta|_1$  是惩罚项,

显示它是岭回归和 Lasso 惩罚项函数的组合。

弹性网回归的回归参数  $\beta$  的估计为  $\hat{\beta} = \arg \min \left[ \sum_{i=1}^n \left( Y_j - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda_1 \sum_{j=1}^p \beta_j^2 + \lambda_2 \sum_{j=1}^p |\beta_j| \right]$ 。

(8) “规则化”方法。是指在 SPSS 的最优尺度回归中, 增加的【规则化】对话框中可供选择的方法, 主要有岭回归法、Lasso 回归法和弹性网回归法。

(9) 符号模式 (模型)。是指对象的组成元素与相互关系都有逻辑符号表示, 是概念模型的一种。

## 5. 关于缺失值的插补

与标准回归分析一样, 对于存在缺失值的变量, 如果不作替换处理, 则要么在最优尺度回归分析中剔除该变量, 要么将缺失值对应的记录整个删除。因此, 为了保证有足够的变量和样本量加入到回归分析中, 对缺失值进行插补是一种不错的选择。

在最优尺度回归分析中, 当含有缺失值的变量  $j$  被指定主动用插补众数或额外的类别处理时, 首先这些变量用  $k_j$  (变量  $j$  的分类数) 来计算要优先于对这些个案进行剔除, 然后用具有最大的加权频数 (众数; 如果有多个众数存在, 则用最小的一个类别) 的分类指示符, 或用  $k_j+1$  (额外的类别) 来估算。

如果在【缺失值】选项卡的【方案】栏中选择了【排除此变量具有缺失值的对象】选项，则使用个案剔除法，那样  $k_j$  将被调整。

如果对变量用样条名义、样条有序、有序或名义最优尺度水平来估算额外的类别，则在最后阶段尺度水平的约束里，不包括额外的类别。

有关样条名义、样条有序、有序或名义最优尺度水平的计算方法，将在下面的分类变量的最优量化处理方法中作进一步介绍。

6. 分类回归中的目标函数、分类变量最优量化处理方法及目标函数最优化

设响应变量的类别量化值的  $k_r$  阶向量用  $y_r$  表示，预测变量  $j$  的类别量化值的  $k_j$  阶向量用  $y_j$  表示，预测变量的回归系数的  $p$  阶向量用  $b$  表示，预测变量的索引集为  $J_p$ ， $b_j$  为  $s_j \times t_j$  阶样条系数向量， $t_j$  为内部节点数。

(1) 目标函数。实际上，分类回归的目的是要寻找一组  $y_r$ 、 $b$  以及  $y_j$ ，在  $j \in J_p$ ，及  $y_r' D_r y_r = n_w$  的条件约束下，以使目标函数

$$\sigma(y_r; b; y_j) = \left( G_r y_r - \int_{j \in J_p} b_j G_j y_j \right) W \left( G_r y_r - \int_{j \in J_p} b_j G_j y_j \right)$$

有最小值。响应变量的量化值同样也被中心化处理，即它们满足  $u' W G_r y_r = 0$ ， $u$  表示元素全为 1 的  $n$  维向量。

在选择【规则化】中的方法之后，受到下述条件约束的损失函数：

$$\begin{aligned} \int_{j \in J_p}^p \beta_j^2 &\leq t_2 \quad \text{适用于岭回归} \\ \int_{j \in J_p}^p |\beta_j| &\leq t_1 \quad \text{适用于 Lasso 回归} \\ \int_{j \in J_p}^p |\beta_j| &\leq t_1 \text{ 和 } \int_{j \in J_p}^p \beta_j^2 \leq t_2 \quad \text{适用于弹性网回归} \end{aligned}$$

式中， $\beta_j$  为预测变量  $j$  的回归系数； $p$  为预测变量的数量。

受到条件约束的损失函数也可以被写成下列的惩罚性损失函数：

$$\begin{aligned} L^{\text{ridge}} &= L + \lambda_2 \int_{j \in J_p}^p \beta_j^2 \\ L^{\text{lasso}} &= L + \lambda_1 \int_{j \in J_p}^p \text{sign}(\beta_j) \beta_j \\ L^{\text{enet}} &= L + \lambda_1 \int_{j \in J_p}^p \text{sign}(\beta_j) \beta_j + \lambda_2 \int_{j \in J_p}^p \beta_j^2 \end{aligned}$$

式中， $\lambda_1$  是套索回归惩罚值； $\lambda_2$  为岭回归惩罚值； $L$  为一般线性回归的损失函数。

(2) 分类变量的最优量化处理方法。在 SPSS 的分类回归中，提供了下述 5 种最优尺度水平，可以用来对分类回归中使用的变量进行最优化定量处理，所有变量可单独地选择这些最优尺度水平。对所有这些选项的基本要求是同等类别的指示符可以得到相等的量化值。

① 名义尺度水平。只有等同约束。

② 样条名义尺度水平。 $y_j = d_j + S_j a_j$  (有等同和样条约束)。式中， $d_j$  为样条截距； $S_j$  为变量  $j$  的  $k_j(s_j \times t_j)$  阶的 I 型样条基函数多项式次数； $t_j$  为内部结点的数量  $a_j$  限定为包含非负元素 (以确保单调 I 型样条)； $y_j$  为变量  $j$  的量化值， $j=1, \dots, m$ 。

③ 样条有序尺度水平。 $y_j = d_j + S_j a_j$  (有等同和单调性样条约束)。

④ 有序尺度水平。 $y_j \in C_j$  (有等同和单调性约束)。单调性约束  $y_j \in C_j$  意指  $y_j$  必须位于非递减元素的所有  $k_j$  维向量的凸锥上。 $k_j$  为变量  $j$  的类别数量 (包括增补对象)。

⑤ 数值尺度水平。 $y_j \in L_j$  (有等同和线性约束)。线性约束  $y_j \in L_j$  意指包含  $k_j$  连续整数的向量的线性转换  $y_j$  必须位于所有  $k_j$  向量的子空间。

为达到鉴别之目的,  $y_j$  总是被标准化处理, 以便使得  $y_j' D_j y_j = n_w$ 。式中,  $n_w$  是加权分析样品的数量, 假设对象  $i$  的权重为  $w_i$ ; 如果对象没有加权, 则  $w_i = 1$ , 如果对象  $i$  是增补对象, 则  $w_i = 0$ 。因此,  $n_w$  可用  $n_w = \sum_{i=1}^n w_i$  来计算得到。另外, 式中的  $D_j$  为  $k_j \times k_j$  对角矩阵,  $D_j = G_j' W G_j$ , 其中,  $W$  是对角线元素为  $w_i$  的  $n_{\text{tot}} \times n_{\text{tot}}$  阶对角矩阵,  $n_{\text{tot}}$  是 (分析+增补) 样品的总数。 $G_j$  是变量  $j$  的  $n_{\text{tot}} \times k_j$  阶指标矩阵,  $G_j$  中的元素被定义为

对于  $i = 1, \dots, n_{\text{tot}}$ ;  $r = 1, \dots, k_j$ ,  $r$  为响应变量的索引, 有

$$g_{(j)ir} = \begin{cases} 1 & \text{当第 } j \text{ 个对象在变量 } j \text{ 的第 } r \text{ 个类别中时} \\ 0 & \text{当第 } j \text{ 个对象不在变量 } j \text{ 的第 } r \text{ 个类别中时} \end{cases}$$

(3) 目标函数最优化。可通过执行迭代计划来实现。这个迭代计划包括:

① 初始化。类别量化的初始化可通过下述方式来完成。

- 随机化: 初始类别量化值  $\tilde{y}_j (j = 1, \dots, m)$  被定义为变量  $j$  的  $k_j$  类指示符的标准化值, 这可得  $u' W G_j \tilde{y}_j = 0$ ,  $y_j' D_j y_j = n_w$ , 且初始回归系数是与响应变量的相关系数。
- 数值化: 在这种情况下, 执行迭代计划两次。在第一次循环中, (用初始化①的值来初始化) 所有变量被当作数值尺度水平处理; 第二次循环中, 用指定的尺度水平开始类别量化, 并且回归系数使用第一次循环中的值。
- 多点搜索 (所有): 当为一个或多个预报因子指定了样条有序水平, 或有序尺度水平时 (Van der Kooij, Meulman, and Heiser, 2006), 选择【多个系统的起始值】中的【所有可能的符号模式】选项, 以确保得到全局最优解。在选择这个选项时, 将执行  $2^s$  次迭代计划, 其中,  $s$  是具有 (样条) 有序尺度水平的预测变量的数量, 并且  $2^s$  是 (样条) 有序尺度水平的预测变量的回归系数的所有可能的符号模式的数量。每次执行迭代计划用相同初始类别量化值和回归系数 (用初始化①的值来初始化) 开始, 但系数有不同的符号模式。在迭代过程中, 固定符号。最后, 使用最优符号模式 (如果使用规则化中的方法, 则该符号模式能导致最大  $R^2$  或决定系数) 执行 1 次或多次迭代计划。
- 多点搜索 (值): 当在【多个系统的起始值】选项中指定阈值时, 选择 (样条) 有序尺度水平的预测变量的回归系数的符号模式则执行两次迭代计划。使用组合的损失方差的百分比策略和分层策略 (Van der Kooij, Meulman, and Heiser, 2006), 选择符号模式。在本选项中符号模式的最大数量为  $1 + \sum_{i=1}^s i$ 。

在第一个周期中 (用初始化①的值来初始化) 所有变量被当作名义水平处理。在第二个周期中, 使用指定的尺度水平, 用第一个周期中获取的类别量化值和回归系数作为它们的起始值。在第二个周期的一次迭代之后, (样条) 有序尺度水平预测变量在第一个周期的最后一次迭代与第二个周期中第一次迭代中的方差的减少被确定。如果预测变量减少的百分比超出指定的阈值, 则允许预测变量为负号。然后第二个周期继续一定

次数的迭代：一次使用所有(样条)有序水平预测变量的正符号的回归系数以及  $q$  次使用一个负符号的(样条)有序水平预测变量的回归系数，这里  $q$  是允许有负号的(样条)有序水平预测变量的数量。如果“所有正”符号模式比“所有负”符号模式给出更好的结果(如果使用【规则化】中的方法，则有更大的  $R^2$  或决定系数)，则执行再一次使用“所有正”符号模式迭代计划。否则，如果“一个负”符号模式之一比“所有负”符号模式给出更好的结果，则选择最好的“一个负”符号模式，并且对“二个负”符号模式(通过向最好的“一个负”符号模式中再增加一个负号模式)重复第二个周期。然后，“二个负”符号模式的结果与“一个负”符号模式作比较，如果“一个负”符号模式的结果更好，则选择“一个负”符号模式。否则，对“三个负”符号模式进行重复第二个周期，以此类推。

- 固定符号：在这种情况下里，执行两次迭代计划。在第一次循环里，(用初始化①的值来初始化)所有变量被当作名义水平处理。第二次循环中，利用指定的尺度水平，用第一次循环中获取的类别量化值和回归系数及(样条)有序尺度水平的预测变量的回归系数固定的符号(在用户指定文件中读取)作为它们的起始值。

② 对响应变量更新类别量化值。用固定的当前值  $y_j$ ， $j \in j_p$ ，无约束的  $y_r$  的更新值为  $\tilde{y}_r = D_r^{-1} G'_r W_r$ 。

名义水平： $y_r^* = \tilde{y}_r$ 。

对于下面 4 个最优尺度水平，如果变量  $j$  是用额外类别来估算的，则  $y_r^*$  在初始阶段里包括类别  $k_r$ 。

- 样条名义水平和样条有序水平： $y_r^* = d_r + S_r a_r$ 。

计算的样条转换值作为在基于  $S_r$  的 I-样条的基础上的  $\tilde{y}_r$  的加权回归 ( $D_r$  的权重对角线元素)。对于受到非负约束的  $a_r$  的样条有序尺度水平的元素，它使得  $y_r^*$  单调增加。

- 有序水平： $y_r^* \leftarrow \text{WMON}(\tilde{y}_r)$ 。符号  $\text{WMON}()$  用来表示加权单调的回归过程，它使得  $y_r^*$  单调增加。使用的权重是  $D_r$  的对角线元素，并且使用的子算法是上下区最小违背算法 (Kruskal, 1964; Barlow et al., 1972)。
- 数值水平： $y_r^* \leftarrow \text{WLIN}(\tilde{y}_r)$ 。符号  $\text{WLIN}()$  用来表示加权线性回归过程。使用的权重是  $D_r$  的对角线元素。

接下来  $y_r^*$  被标准化处理(如果响应变量使用额外类别来估算，则  $y_r^*$  从此以后包括类别  $k_r$ )： $y_r^+ = n_w^{1/2} + y_r^* (y_r'^* D_r y_r^*)^{-1/2}$ 。

③ 对预测变量更新类别量化值和回归系数。为更新预测变量  $j$ ， $j \in j_p$ ，首先变量  $j$  的贡献从  $v: v_j = v - b_j G_j y_j$  中消去。然后，无约束的  $y_r$  的更新值为  $\tilde{y}_j = D_j^{-1} G'_j W (W_r y_r - v_j)$ ，接下来按步骤②中的要求对  $\tilde{y}_j$  进行约束及标准化处理，从而得到  $y_j^+$ 。

最后，更新回归系数

$$b_j^+ = n_w^{-1} \tilde{y}_j' D_j y_j^+$$

得到如下的正规化回归系数

$$\beta_j^+ = \frac{\beta_j^*}{1 + \lambda_2}$$

适用于邻回归。



如果  $\beta_j^* > 0$ , 则  $\beta_j^+ = \left(\beta_j^* - \frac{\lambda_1}{2} w_j\right)_+ = \beta_j^* - \frac{\lambda_1}{2}$ , 如果  $\beta_j^* < 0$ , 则  $\beta_j^+ = \beta_j^* + \frac{\lambda_1}{2}$ , 适用于 Lasso

回归。

并且, 如果  $\beta_j^* > 0$ , 则  $\beta_j^+ = \frac{\left(\beta_j^* - \frac{\lambda_1}{2} w_j\right)_+}{1 + \lambda_2} = \frac{\left(\beta_j^* - \frac{\lambda_1}{2}\right)}{1 + \lambda_2}$ , 如果  $\beta_j^* < 0$ , 则  $\beta_j^+ = \frac{\left(\beta_j^* + \frac{\lambda_1}{2}\right)_+}{1 + \lambda_2}$ ,

适用于弹性网回归 (van der Kooij, 2007)。

④ **收敛检验**。用连续的显性预测误差值 (APE) 之间的差值与用户指定收敛性判定标准  $\varepsilon$  (一个小正数) 进行比较。

APE 可用下式计算:

$$\text{APE} = n_w^{-1} \left( \mathbf{G}_r \mathbf{y}_r - \int_{j \in J(p)} \boldsymbol{\beta}_j \mathbf{G}_j \mathbf{y}_j \right)' \mathbf{W} \left( \mathbf{G}_r \mathbf{y}_r - \int_{j \in J(p)} \boldsymbol{\beta}_j \mathbf{G}_j \mathbf{y}_j \right)$$

在没有选择【规则化】中的方法时, APE 等于 1 减去多元回归系数的平方。只要 APE 差值超过  $\varepsilon$  则重复步骤②~③。

## 7. 选择规则化方法

如果指定正规化使用【规则化】中的方法, 则所有上述诊断方法也同样应用于选择或指定的【规则化】中的模型。如果指定的模型不止一个 (不止一个惩罚值), 那么可以要求对每个模型进行诊断。

(1) 标准的回归系数和用下式计算。

① 适用于岭回归的标准的回归系数和:  $\frac{\int_{j \in J_p}^p \beta_j^2}{\int_{j \in J_p}^p (\beta_j^*)^2}$ 。

② 适用于 Lasso 回归和弹性网回归的标准的回归系数和:  $\frac{\int_{j \in J_p}^p \text{sing}(\beta_j) \beta_j}{\int_{j \in J_p}^p \text{sing}(\beta_j^*) \beta_j^*}$ 。

(2) **显性预测误差 (APE)**。在最优化算法最后迭代的收敛步骤中计算得到 APE。

(3) **期望预测误差 (EPE)**。为标准的 (量化) 的数据计算期望预测误差。仅当为所有变量指定数值尺度水平时, 同样为原始数据计算 EPE。

① **增补对象 (测试样品)**。

• 训练数据 (活动样品) 的期望预测误差为

$$\text{EPE}^{\text{train}} = \frac{1}{n_w} \sum_{i=1}^n \left[ (\mathbf{G}_r \mathbf{y}_r)_i - \left( \sum_{j \in J_p} \boldsymbol{\beta}_j \mathbf{G}_j \mathbf{y}_j \right)_i \right]^2$$

并且标准误为

$$\text{SE}^{\text{train}} = \left[ \frac{1}{n_w^2} \sum_{i=1}^n w_i (\text{EPE}_i^{\text{train}} - \text{EPE}^{\text{train}})^2 \right]^{1/2}$$

• 检验数据 (增补对象) 的期望的预测误差为

$$\text{EPE}^{\text{test}} = \frac{1}{n_{\text{tot}} - n} \sum_{i \in S} \left[ (\mathbf{G}_r \mathbf{y}_r)_i - \left( \sum_{j \in J_p} \boldsymbol{\beta}_j \mathbf{G}_j \mathbf{y}_j \right)_i \right]^2$$

式中,  $s$  是增补对象的索引集。其标准误为

$$\text{SE}^{\text{test}} = \left[ \frac{1}{(n_{\text{tot}} - n)^2} \sum_{j \in S} (\text{EPE}_i^{\text{test}} - \text{EPE}^{\text{test}})^2 \right]^{1/2}$$

- 对于增补类别的量化的估计值(增补样品中只出现一个类别), 见下面的量化部分。

用  $\text{EPE}^{\text{train}}$ 、 $\text{SE}^{\text{train}}$ 、 $\text{EPE}^{\text{test}}$ 、 $\text{SE}^{\text{test}}$  乘以

$$\frac{1}{n_w} \sum_{i=1}^n \left( h_{ri} - \frac{1}{n_w} \sum_{i=1}^n h_{ri} \right)^2$$

(每个活动样品响应变量的方差)则为原始数据产生 EPE 和 SE。

② **重复采样, 0.632 自举法(Bootstrap)**。通过从活动对象(训练数据)中随机(有放回)抽取  $n$  次来建立自举法的数据集, 包括对象(样品)权重。自举法中的 EPE 可用下式计算:

$$\text{EPE}^{\text{boot}} = \widehat{\text{Err}}^{(0.632)} = \overline{\text{err}} + \widehat{\text{OP}}$$

式中, OP(乐观值)用下式估计:

$$\widehat{\text{OP}} = 0.632(\overline{\text{Err}}^{(1)} - \overline{\text{err}})$$

并且  $\overline{\text{Err}}^{(1)}$ , 预测误差的留一法自举估计为

$$\overline{\text{Err}}^{(1)} = \frac{1}{n_w^{(1)}} \sum_{i=1}^n \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} w_i \left[ (\mathbf{G}_r \mathbf{y}_r^b)_i - \left( \sum_{j \in J_p} \boldsymbol{\beta}_j^b \mathbf{G}_j \mathbf{y}_j^b \right)_i \right]^2$$

适用于  $|C^{-i}| \neq 0$ 。式中,  $C^{-i}$  是自举样本  $b$  ( $b=1, \dots, B$ ) 的索引集, 它

- 不包括观察  $i$ ;
- 对于名义或有序水平转换的变量, 包括应用于观察  $i$  的类别;
- 对于样条转换变量的观察  $i$  不需要使用外推法。

$n_w^{(1)}$  是  $|C^{-i}| \neq 0$  的观察的数量。(集合  $|C^{-i}|$  可以为空集, 例如, 假如观察  $i$  在样条转换的变量上只有一个极端的类别, 而且这个类别的频数为 1, 则每个自举样本不包括这个观察, 也不包括极端的类别, 因此适用于观察  $i$  的所有自举样本被排除。)

自举法中的标准误用下式计算:

$$\text{SE}^{\text{boot}} = \left[ \frac{1}{n_w^2} \sum_{i=1}^n w_i (\overline{\text{Err}}_i^{(1)} - \overline{\text{Err}}^{(1)})^2 \right]^{1/2}$$

自举样本  $b$  里的样品在计算  $\overline{\text{Err}}^{(1)}$  中添加乘以响应变量的方差  $[\dots, w_i \text{var}(\mathbf{h}_r^b)(\dots)]$ , 则得到原始数据的 EPE 和 SE。

③ **重复采样, 交叉验证**。数据被随机地分到活动对象(训练数据)的  $k$  个不相交子集中, 包括对象(样品)权重。重复采样中的 EPE 可用下式计算:

$$\text{EPE}^{\text{CV}} = \frac{1}{n_w} \sum_{i=1}^n \sum_{j \in k} w_i \left[ (\mathbf{G}_r \mathbf{y}_r^k)_i - \left( \sum_{j \in J_p} \boldsymbol{\beta}_j^{-k} \mathbf{G}_j \mathbf{y}_j^{-k} \right)_i \right]^2$$

式中,  $k(k=1, \dots, K)$  索引第  $k$  个子集,  $-k$  索引其余的数据部分。

其标准误可用下式计算:

$$SE^{CV} = \left[ \frac{1}{n_w^2} \sum_{i=1}^n w_i (EPE_i^{CV} - EPE^{CV})^2 \right]^{1/2}$$

删除第  $k$  部分的样品, 在计算  $EPE^{CV}$  时, 只要加入乘以响应变量的方差  $[\dots, w_i \text{var}(\mathbf{h}_r^{-k})(\dots)]$ , 则可得到原始数据的 EPE 和 SE。

在自举样本中或在第  $k$  个部分删除的数据中不会发生类别的量化, 可按增补类别估计。

## 8. 相关系数

在转换前, 变量之间的相关系数  $R$  可用下式计算:

$$\mathbf{R} = n_w^{-1} \mathbf{H}_c' \mathbf{W} \mathbf{H}_c$$

式中,  $\mathbf{H}_c$  为加权的中心化, 并且标准化的  $\mathbf{H}$  中不包括响应变量。

在转换后, 变量之间的相关系数  $\mathbf{R}$  可用下式计算:

$$\mathbf{R} = n_w^{-1} \mathbf{Q}' \mathbf{W} \mathbf{Q}$$

式中,  $\mathbf{Q}$  的列向量为  $\mathbf{q}_j = \mathbf{G}_j \mathbf{y}_j$ ,  $j \in J_p$ 。

(1) **0 阶相关系数**。在转换的响应变量  $\mathbf{G}_r \mathbf{y}_r$  和转换的预测变量  $\mathbf{G}_j \mathbf{y}_j$  之间的相关系数为

$$r_{ij} = n_w^{-1} (\mathbf{G}_r \mathbf{y}_r)' \mathbf{W} \mathbf{G}_j \mathbf{y}_j$$

(2) **偏相关系数**为

$$\text{偏相关系数}_j = b_j [(1/t_j)(1-R^2) + b_j^2]^{-1/2}$$

式中,  $t_j$  为变量  $j$  的容忍度。

在使用正规化方法时, 按下式计算常规最小二乘法的回归系数:

$$\boldsymbol{\beta}^* = (n_w \mathbf{R})^{-1} \mathbf{Q}' \mathbf{W} (\mathbf{G}_r \mathbf{y}_r)$$

使用  $R_p$  的特征值和特征向量计算转换后的相关矩阵  $\mathbf{R}$  和  $\mathbf{R}^{-1}$ 。其中,  $\mathbf{R}_p$  是回归系数大于 0 的预测变量的相关矩阵, 并且  $R^2$  用下式来计算:

$$R^2 = \{(\mathbf{G}_r \mathbf{y}_r)' \mathbf{W} \mathbf{Q} \boldsymbol{\beta}^* [n_w (\mathbf{Q} \boldsymbol{\beta}^*)' \mathbf{W} \mathbf{Q} \boldsymbol{\beta}^*]^{-1/2}\}^2$$

(3) **部分相关系数**为

$$\text{PartCorr}_j = b_j t_j^{1/2}$$

式中,  $t_j$  为变量  $j$  的容忍度。

## 9. 标准回归系数 $\beta$ 及其标准误

(1) 标准回归系数可用  $\beta = b_j$  来获取。

(2)  $\beta$  的标准误用下式计算:

$$SE(\text{Beta}) = [(1-R^2) / (n_w - l - \mathbf{u}' \mathbf{f}) t_j]^{1/2}$$

式中,  $t_j$  为变量  $j$  的容忍度。

## 10. 自由度

一个变量的自由度取决于最优尺度水平:

① 数值水平变量的自由度:  $f_j = 1$ 。

② 样条有序、样条名义水平变量的自由度:  $f_j = s_j + t_j$  减去  $a_j$  中元素等于 0 的个数。

③ 有序、名义水平变量的自由度:  $f_j = y_j$  中不同值的数量减 1。

11. 重要性

相对重要性的 Pratt 测度(Pratt, 1987), 有

$$\text{Im } p_j = b_j r_{jj} / R^2$$

仅当使用非正规化处理时显示相对重要性。

12. 容忍度

最优尺度预测变量的容忍度用下式计算:

$$t_j = r_{pjj}^{-1}$$

式中,  $r_{pjj}^{-1}$  是  $R_p$  的第  $j$  个对角元素,  $R_p$  是回归系数大于 0 的预测变量的相关矩阵。

用同样的方式, 如果适用, 则利用离散化、插补法及删除法, 使用原始预测变量的相关矩阵计算并报告原始预测变量的容忍度。

11.10.2 最优尺度回归过程

- (1) 按【分析→回归→最优尺度】顺序打开对话框, 见图 11-79。
- (2) 从左侧的源变量框中选择一个分类变量作为因变量进入【因变量】框中。
- (3) 单击因变量框下的【定义度量】(应为定义尺度)按钮, 打开如图 11-80 所示的【类别回归: 定义度量】对话框, 设置因变量的最优尺度水平。默认情况下, 它们采用有两个内部节点的二次单调性样条有序水平。此外, 还可以设置分析变量的权重。



图 11-79 【分类回归】对话框

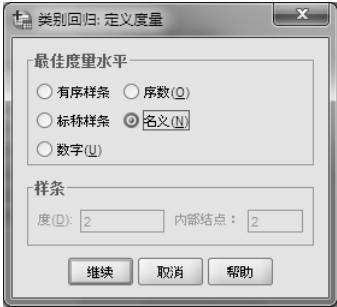


图 11-80 【类别回归: 定义度量】对话框

- ① 【最佳度量水平】(应为最佳尺度水平)栏。提供了 5 种可以用于量化每个变量的尺度水平。
  - 【有序样条】。变换后的最优尺度变量中保留观测变量的分类顺序。类别点将位于一条通过原点的直线上(矢量)。转换结果是一个选定次数的平滑单调的分段多项式。样条的每一段都是按用户指定的次数并按内部结点的确定位置生成的。
  - 【标称样条】(名义样条)。变换后的最优尺度变量中只保留观测变量的信息是按分类构成的对象分组, 而观测变量的分类顺序则不再保留。类别点将位于一条通过原点的直线上(矢量)。转换结果是一个预先选定次数的平滑的、可能非单调的分段多项式。样条的每一段都是按用户指定的次数并按内部结点的确定位置生成的。
  - 【数字】。分类将被视为有序且等间距。变换后的最优尺度变量中保留观测变量的分类

号之间的分类顺序和等间距性。类别点将位于一条通过原点的直线上(矢量)。当所有变量都为定量变量时,该分析类似于主成分分析。

- **【序数】**。变换后的最优尺度变量中保留观测变量的分类顺序。类别点将位于一条通过原点的直线上(矢量)。转换结果比有序样条转换拟合得好,但是平滑度较低。
- **【名义】**。变换后的最优尺度变量中只保留观测变量的信息是按分类构成的对象分组,而观测变量的分类顺序则不再保留。类别点将位于一条通过原点的直线上(矢量)。转换结果比名义样条转换拟合得好,但是平滑度较低。

② **【样条】**栏。当选择**【有序样条】**或**【标称样条】**选项时,对样条函数的次数和节点个数进行定义。

- **【度】**(应为次数)。样条函数的次数,系统默认值为 2。
- **【内部结点】**。内部结点数,系统默认值为 2。

(4) 从源变量框中选择一个或多个自变量进入**【自变量】**框中。单击**【定义度量】**按钮,在打开的对话框中对自变量的最优尺度水平进行设置。

(5) 单击**【离散化】**按钮,打开如图 11-81 所示的对话框,可以选择对原变量重新进行编码的方法。如果在该对话框中没有作任何选择,那么系统对有小数值的变量将分成具有近似正态分布的 7 个类别(如果变量的不同值的数目小于 7,则将按此数目划分类别),按字母数值升序顺序分配类别指示符。字符串变量总是转换为正整数,这些整数可用来对字符串变量进行离散化处理。默认情况下,其他变量保留原样。随后的分析中将使用离散化变量。

在**【变量】**框中,系统自动列出方程中的因变量和全部自变量,并在变量名后附有(未指定)。

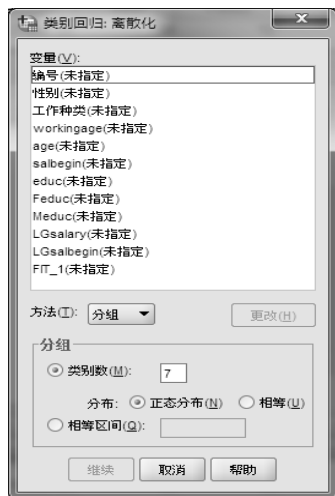


图 11-81 **【类别回归: 离散化】**对话框

① **【方法】**下拉列表。可以选择一种离散化并重新编码的方法。

- **【未指定】**采用系统默认形式。
- **【分组】**按指定的分类数或按等间距进行重新编码。
- **【秩】**用样品排序后的秩来代替原变量的值。
- **【乘】**对原变量进行标准化处理。然后将标准化的值乘以 10 并进行四舍五入处理,再加上一个值使得其最小值为 1。

如果需要对某个变量重新编码,则先选中该变量,再选择重新编码方法,然后单击右侧的**【更改】**按钮即可。完成更改的变量,在其变量名后将显示出相应更改的方法,同时,**【继续】**按钮将被激活。

② **【分组】**栏。仅当在**【方法】**下拉列表中选择用**【分组】**选项对变量进行离散化处理时,方可使用以下选项:

- **【类别数】**(分类数)框指定分类的数量,并指定这些类别间的变量值是服从近似**【正态分布】**,还是服从**【均匀分布】**(选项汉化为**【相等】**,不太确切)。系统默认分类数的值为 7。
- **【相等区间】**(等间隔)框。必须输入相应的数值来设定间隔的长度。系统将按设定间隔的大小对原变量进行分类。

(6) 单击【类别回归：缺失值】按钮，打开如图 11-82 所示的对话框，可以选择处理分析变量及其缺失值的补充方案。

①【缺失值方案】栏。只有【分析变量】框，它列出了因变量和所有的自变量，并在所有变量名后附有处理缺失值的默认方式的标示【排除】。

②【方案】栏。

- 【排除此变量具有缺失值的对象】。对于在【分析变量】框中选定的分析变量而言，分析时将剔除具有缺失值的对象。它是系统默认选项，此方案对补充变量不适用。
- 【为缺失值归因】。

A.【众数】。用众数所在组的类别值来替代缺失值，如果有多个众数，则用其中的最小一个类别值来替代缺失值。

B.【附加类别】。将缺失值替换为相同的一个额外划分的类别值。这就意味着该变量中有缺失值的对象被视为属于同一个(附加)类别。

选择完成后，单击【分析变量】框右下角的【更改】按钮，提交系统运行。

(7) 单击【选项】按钮，打开如图 11-83 所示的【类别回归：选项】对话框，可以指定迭代和收敛条件，选择补充对象和设置绘图标记。



图 11-82 【类别回归：缺失值】对话框

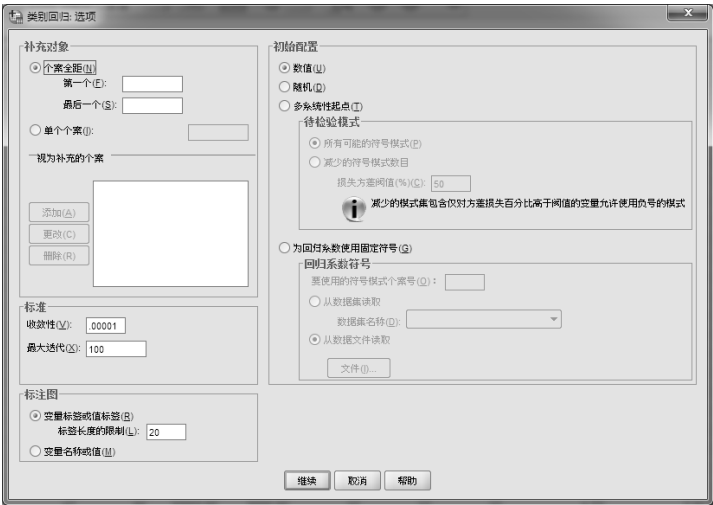


图 11-83 【类别回归：选项】对话框

①【补充对象】(增补对象)栏。可以指定要视作补充对象的对象。如果认为数据文件中的某些记录不太可靠或不太重要，可对其作标示，将其视作补充对象。

- 【个案全距】(个案范围)选项。如果有一个连续区域的记录被视作补充对象，则选择本项。在【第一个】及【最后一个】框中输入这些记录的起、止序号，可指定个案范围。
- 【单个个案】。在其后框中输入被视为补充对象的序号。指定后，单击下面的【添加】按钮，则可将选定的个案添加到【视为补充的个案】框中。

选定为补充对象后,不可以对其作加权处理,也即原先对其指定的权重将无效。

## ②【初始配置】栏。

- **【数值】**。系统默认选项。当所有变量均为尺度或有序测度的变量时,选择此项。
- **【随机】**。如果变量中至少有一个变量是名义测度变量,则选择此项。
- **【多系统性起点】**(多个系统的起始值)。如果至少有一个变量在量化过程中使用有序或有序样条尺度水平,则通常的模拟拟合算法可能得到的解欠佳。在此情况下,可以选择本项。本项中还有两种选择:

A. **【所有可能的符号模式】**(应为符号模型)。由于其具有**【多系统性起点】**的所有可能的待检验的符号模型,因此可以始终寻找最优解。但由于数据集中的有序和有序样条变量的数量增加,因此所需的处理时间也大大增加。

B. **【减少的符号模式数量】**。为提高运算速度,可以选择本项。可以在**【损失方差阈值(%)】**框中指定方差损失阈值百分比来减少检验的符号模型的数量。阈值越高,排除的符号模型越多。在选用此项时,尽管无法保证获得最优解,但也消除了得到的解欠佳型的可能性。此外,如果找不到最优解,欠佳的解与最优解也不会差别太大。在选定**【多系统性起点】**时,每个起点的回归系数符号被写入外部 SPSS Statistics 数据文件或当前工作的数据集中。

- **【为回归系数使用固定符号】**。如果要采用先前**【多系统性起点】**的运行结果,则可选择本项。须在此选项下的**【回归系数】**栏的**【要使用的符号模式个案号】**框中输入需要用到本处理的个案序号。

为回归系数使用的固定符号可在指定的数据集或数据文件中读取。符号采用“1”和“-1”进行标示,它需要在指定数据集或文件的某行中。如果此前已经对本数据文件作过分类回归处理,并为回归系数符号创建了数据集或数据文件,则下面两个选项被激活:

- A. **【从数据集读取】**。在数据集名称后的框中输入所需的数据集名称。
- B. **【从数据文件读取】**。单击**【文件】**按钮,在文件浏览器中选定所需的文件。

## ③【标准】栏。可指定回归计算中执行的最大迭代次数和拟合的收敛标准。

- **【收敛性】**框。输入拟合的标准。系统默认值为 0.00001。
- **【最大迭代】**框。输入最大的迭代次数。系统默认值为 100。

如果上两次迭代之间的总拟合之差小于收敛值,或者达到了最大迭代次数,则回归的迭代过程终止。

④**【标注图】**栏。可用来指定在图中使用变量标签(或变量值标签)或变量名(或变量值)及标签的最大长度。

- **【变量名标签或值标签】**。需在**【标签长度的限制】**框中输入定义标签的最大长度的值。系统默认值为 20。
- **【变量名称或值】**。在输出的图中使用变量名或变量值。

(8) 单击**【规则化】**按钮,打开如图 11-84 所示的**【类别回归: 规则化】**对话框,可以设定规则化的方法及产生弹性网图。

- ①**【方法】**栏。规则化方法可以使回归系数估计缩小为 0,以降低其变异性,从而改善模型



图 11-84 【类别回归: 规则化】对话框

的预测误差。在选取了规则化方法后,每个惩罚系数值的规则化模型和系数被写入外部 SPSS Statistics 数据文件或当前工作的数据集中。

- **【无】**。不施加惩罚系数的约束。
- **【Ridge 回归】**(岭回归)法。在回归过程中,引入惩罚项以缩小系数,惩罚项等于系数平方乘以惩罚系数的总和。该系数可从 0(无惩罚)到 1 变化;如果指定了范围与增量,那么过程将寻求“最佳”的惩罚值。
- **【套索】**(应为 Lasso)。Lasso 的惩罚项是建立在绝对系数总和的基础上的,惩罚系数的指定与岭回归类似,但 Lasso 涉及的计算量更大。
- **【弹性网络】**法。“弹性网络”简单地将 Lasso 和岭回归惩罚两者结合在一起,在给定的网格中搜寻以发现“最佳”的 Lasso 和岭回归惩罚系数。对于给定的 Lasso 与 Ridge 回归惩罚,“弹性网络”的计算量并不比 Lasso 多很多。

② **【弹性网络图】**栏。如果选择了**【弹性网络】**法,则按岭回归的惩罚值生成各自的规则化图。

- **【产生所有可能的弹性网络图】**。将由指定的最小和最大 Ridge 回归惩罚值所确定范围内的每个值产生弹性网络图。
- **【为部分 Ridge 回归惩罚产生弹性网络图】**。如果只需产生部分的弹性网络图,则选择本选项,需进一步指定下面两个选项之一:

A. **【值范围】**。需在**【第一个】**和**【最后一个】**框中输入岭回归惩罚值的最小值和最大值。由此确定范围内的值子集的弹性网络图将被在输出窗中显示。

B. **【单值】**。只需在框中输入惩罚值的序号即可,则该值对应的弹性网络图将在输出窗口中显示。

完成上述设定后,单击 Ridge 回归惩罚值下面的**【添加】**按钮,则在 Ridge 回归惩罚值列表中出现所选的结果。

③ **【显示规则化图】**。在输出窗口中输出回归系数与规则化惩罚图。在搜寻某个值范围以寻找“最佳”惩罚系数时,它提供了有关回归系数在该范围上如何变化的视图。

(9) 单击**【输出】**按钮,打开如图 11-85 所示的**【类别回归: 输出】**对话框。在此对话框中,可以选定显示在输出窗口中的统计量。

① **【表】**栏。

- **【复 R】**。在输出窗口中显示  $R^2$ 、调整后的  $R^2$  以及将最优尺度考虑在内的调整后的  $R^2$ 。
- **【ANOVA】**。在输出窗口中显示回归及残差平方和、均方和  $F$  值。它将显示两张 ANOVA 表: 一张表的回归自由度等于预测变量数, 另一张表的回归自由度则是将最优尺度考虑在内的自由度。
- **【系数】**。在输出窗口中显示 3 张系数表:

第 1 张表为方程系数表,包括 beta、beta 的标准误、 $t$  值和  $p$  值。

第 2 张表为系数最优尺度表,包括具有最优尺度的标准化系数  $\beta$  的标准误和自由度。

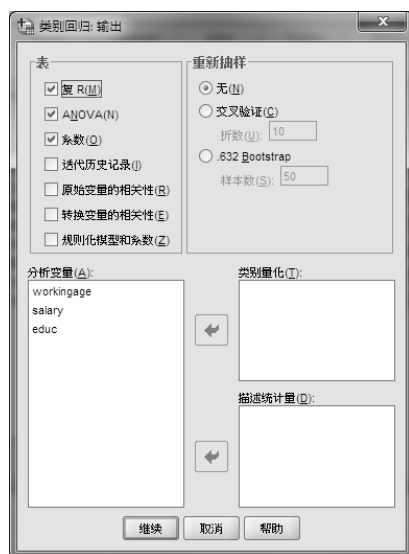


图 11-85 【类别回归: 输出】对话框



第 3 张表为相关系数表, 包括每个变量与因变量的零阶相关系数、部分相关系数和偏相关系数、转换后预测值的相对重要测度以及转换前后的容忍度。

- **【迭代历史记录】**。在输出窗口中显示迭代过程, 包括每次迭代的初始值、复相关系数  $R$  和回归误差。另外, 还列出从第一次迭代开始的复相关系数  $R$  的各次增量。
- **【初始变量的相关性】**。在输出窗口中显示原始变量之间的相关系数矩阵表。
- **【转换变量的相关性】**。在输出窗口中显示转换后变量之间的相关系数矩阵表。
- **【规则化模型和系数】**。在输出窗口中显示每个规则化模型的惩罚值、 $R^2$  和回归系数。如果指定了重新抽样方法, 或指定了补充对象(单个个案), 则在输出窗口中还显示预测误差或均方误差。

② **【重新抽样】** 栏。选择有关模型预测误差的估计方法。

- **【无】**。直接用建模样本估计模型的预测误差。
- **【交叉验证】** 法。交叉验证法将初始样本分割成  $K$  个子样本或群。每一个单独的子样本分别被作为验证样本, 验证样本以外的数据用来训练, 生成分类回归模型, 这样可以得到  $K$  个分类回归模型。对于每个模型, 用其对应的验证样本来估计模型的预测误差。
- **【.632 Bootstrap】** 法。即自举法, 采用有放回方式从数据中随机抽取观察值, 多次重复该过程则获得大量 Bootstrap 样本。为所有 Bootstrap 样本拟合模型, 然后将该拟合模型所估计的模型预测误差应用到非 Bootstrap 样本的个案上。

③ **【分析变量】** 框。提供了可供选择的参与分析的所有变量名。

④ **【类别量化】**。显示选定的变量转换前后值的对照表。

⑤ **【描述统计】**。显示选定变量的描述统计表, 包括频数、缺失值和众数。

(10) 单击 **【保存】** 按钮, 打开如图 11-86 所示的 **【类别回归: 保存】** 对话框, 可以将预测值、残差和转换后的值保存到当前工作的数据集和/或将离散化数据、转换后的值、规则化模型和系数以及回归系数符号保存到外部 SPSS Statistics 数据文件或当前工作的数据集中。



图 11-86 **【类别回归: 保存】** 对话框

在 **【保存】** 对话框中共有两个选项和两个栏目:

① **【将预测值保存到活动数据集】**。在当前活动数据集中自动添加新变量用来存放预测值, 除非在关闭 SPSS 数据编辑器前将活动数据集做过保存处理, 否则再次调用的原数据文件中将不保留这个预测值信息。

②【将残差保存到活动数据集】。将残差保存在当前的数据集中。

③【离散化数据】栏。可指定被离散化处理的数据保存在哪个文件中。

•【创建离散化数据】。在下面单选项中继续进行选择:

A.【创建新数据集】。在【数据集名称】框中输入数据集名。

B.【写入新数据文件】。在将被离散化处理的数据保存到新的数据文件中。单击【浏览】按钮,则在弹出的浏览窗口中选择存放路径,在【文件名】框中输入文件名,单击【保存】按钮,返回到图 11-86 所示对话框。

④【已转换的变量】栏。可指定转换后的变量保存到哪个文件中。

•【将已转换的变量保存到活动数据集】。将转换后的变量保存到当前工作的数据集中。

•【将已转换的变量保存到新数据集或文件】。将转换后的变量保存到新数据集或文件中。

具体操作方法同【离散化数据】栏中的【写入新数据文件】选项。

需要注意的是,保存离散化数据与保存转换变量时,保存的文件名或数据集名应各不相同。

⑤【规则化模型和系数】栏。只要在【类别回归:规则化】对话框中选择了规则化的【方法】,就可以在本栏中保存规则化模型和系数。默认情况下,该过程以唯一名称创建新数据集,当然用户也可以自行指定名称,或将其写入外部文件。

该栏下面有两个选项:【创建新数据集】及【写新数据文件】。具体操作方法同【离散化数据】栏中的【创建新数据集】与【写入新数据文件】选项。

⑥【回归系数符号】栏。只要在【类别回归:选项】对话框上使用【多系统性起点】作为初始配置,就可以在本栏中保存回归系数符号。默认情况下,该过程以唯一名称创建新数据集,当然用户也可以自行指定名称,或将其写入外部文件。

该栏下面有两个选项:【创建新数据集】及【写新数据文件】。具体操作方法同【离散化数据】栏中的【创建新数据集】与【写入新数据文件】选项。



图 11-87 【类别回归:图】对话框

(11) 单击【绘制】按钮,打开如图 11-87 所示的【类别回归:图】对话框,可以提供需要绘制的图形,可以指定将生成转换图和残差图的变量。

在左侧框中列有【类别回归】对话框中选定的因变量和自变量名称。

①【转换图】框。选择左侧框中所要用来作转换图的变量,将其移入本框。对于选定的这些变量,在输出窗中显示每个原始分类变量和定量化后的分类的转换图。空类别出现在水平轴上,但不影响计算。这些空类别通过连接定量化的线中的断点来识别。

②【残差图】框。选择左侧框中所要用来作残差图的变量,将其移入本框。对于选定的这些变量,在输出窗中显示所选变量的残差图。这里的残差是从所有自变量中排除所选变量之后根据因变量的预测值计算的结果,而且最优分类的量化是用分类指示符乘以  $\beta$  得到的。

(12) 单击【确定】按钮提交运算。

### 11.10.3 最优尺度回归分析实例

【例 13】用数据文件 data11-11(1991 年美国社会情况调查)中的 life(生活状况)、regin(地区)、race(种族)、occocat80(职业类型)为自变量,对因变量 Happy(幸福感)进行最优尺度回归分析。

在 SPSS 中打开数据文件 data11-11 后的操作步骤如下：

(1) 按【分析→回归→最优尺度】顺序打开【类别回归】对话框。

(2) 定义因变量及其最优尺度水平。在左侧的源变量框中，选择 Happy 为因变量送入【因变量】框，单击【定义度量】按钮，打开【类别回归：定义度量】对话框，选择【有序样条】选项，单击【继续】按钮，返回【类别回归】对话框。

(3) 定义自变量及其最优尺度水平。在左侧的源变量框中，选择 life、regin、race、occcat80 为自变量送入【自变量】框。在【自变量】框中，选定变量 life，打开【类别回归：定义度量】对话框，选择【有序样条】选项，单击【继续】按钮，返回【类别回归】对话框。在【自变量】框中，一次选定变量 regin、race、occcat80，打开【类别回归：定义度量】对话框，选择【名义】选项，单击【继续】按钮，返回【类别回归：最优尺度】对话框。

(4) 定义选项。单击【选项】按钮，打开【类别回归：选项】对话框。由于变量 regin、race、occcat80 被定义为名义测度，因此，在【初始配置】栏中，选择【随机】选项。

(5) 定义输出。单击【输出】按钮，打开【类别回归：输出】对话框。选择【复 R】、【系数】、【ANOVA】选项。单击【继续】按钮，返回【类别回归】对话框。

(6) 单击【绘制】按钮，打开【类别回归：图】对话框。在左侧框中选择 Happy 变量，将其移入【转换图】框中。单击【继续】按钮，返回【类别回归】对话框。

其他选用系统默认方式。

(7) 单击【确定】按钮，提交运算。输出结果见表 11-81～表 11-85。

表 11-81 个案处理汇总

案例处理汇总	
有效的活动案例	900
a	617
补充案例	0
总计	1517
分析中使用的案例	900

已排除的案例（显示前 30 个）：3 4 6 7 8 9 12 13 16  
17 19 22 24 25 27 29 31  
32 33 34 36 39 42 44 45  
47 51 52 59 60.

表 11-81 显示了参与最优尺度回归中个案的基本情况。样本容量为 1517，其中因有缺失值或异常值的无效个案数为 617，它们不参与回归分析。在表的下方显示了前 30 个被排除个案所在的记录号。

例如，“3”表示在数据文件的第三个记录(行)中有不符合定义分类的值存在，检验原始数据可发现，该值出现在记录 3 的 life 变量中，见图 11-88。该值为 0，不符合分类变量值最小为 1 的规定，从该数据文件的“变量视图”的“缺失”列中可知，0 为缺失值标记。而记录 6、7 的 occcat80 变量的值为缺失值，其余可类推。

	sex	race	region	happy	life	sibs	childs	age	educ	paeduc	maeduc	speduc	prestg80	occcat80
1	2	1	1.00	1	1	1	2	61	12	97	12	97	22	3.00
2	2	1	1.00	2	1	2	1	32	20	20	18	20	75	1.00
3	1	1	1.00	1	0	2	1	35	20	16	14	17	59	1.00
4	2	1	1.00	9	2	2	0	26	20	20	20	97	48	1.00
5	2	2	1.00	2	1	4	0	25	12	98	98	97	42	3.00
6	1	2	1.00	2	0	7	5	59	10	8	6	97	0	.
7	1	2	1.00	1	1	7	3	46	10	8	98	97	0	.
8	2	2	1.00	2	0	7	4	99	16	5	6	97	60	2.00
9	2	2	1.00	2	2	7	3	57	10	6	5	97	0	.
10	2	1	1.00	2	1	1	2	64	14	8	12	20	38	6.00
11	1	1	1.00	2	1	6	0	72	9	12	98	97	36	6.00

图 11-88 被排除的记录

表 11-82 所示为最优尺度回归中的复相关系数(表中为“多 R”，不正确)、判定系数  $R^2$ 、调整  $R^2$ 、回归方程的预测误差。从  $R^2 = 0.153$  可见，回归方程的拟合效果不很理想。

表 11-83 所示为最优尺度回归模型的方差分析表。由表可见， $p = 0.000$ ，小于 0.05，表明所建模型有统计学上的显著性意义。

表 11-82 相关系数统计量表

模型汇总			
多 R	R 方	调整 R 方	明显预测误差
.392	.153	.142	.847

因变量:General Happiness  
预测变量: Is Life Exciting or Dull Region of the United States Race of Respondent Occupational Category

表 11-83 方差分析表

ANOVA					
	平方和	df	均方	F	Sig.
回归	138.101	12	11.508	13.398	.000
残差	761.899	887	.859		
总计	900.000	899			

因变量:General Happiness  
预测变量: Is Life Exciting or Dull Region of the United States Race of Respondent Occupational Category

表 11-84 回归系数表

系数					
	标准系数		df	F	Sig.
	Beta	标准误差的 Bootstrap			
		(1000) 估计			
Is Life Exciting or Dull	.371	.043	3	74.301	.000
Region of the United States	.030	.025	2	1.434	.239
Race of Respondent	.103	.038	2	7.411	.001
Occupational Category	.063	.027	5	5.554	.000

因变量: General Happiness

表 11-85 相关性和容忍度量表

	相关性			重要性	容忍度	
	零阶	偏	部分		转换后	转换前
Is Life Exciting or Dull Region of the United States	.372	.373	.370	.900	.993	.973
Race of Respondent	-.017	.032	.030	-.003	.967	.995
Occupational Category	.101	.109	.101	.068	.972	.993
	.086	.068	.063	.035	.987	.968

因变量: General Happiness

根据表 11-84 中得到的标准回归系数，可知最优尺度回归方程为

happy = 0.371life + 0.030reign + 0.103race + 0.063occcat80

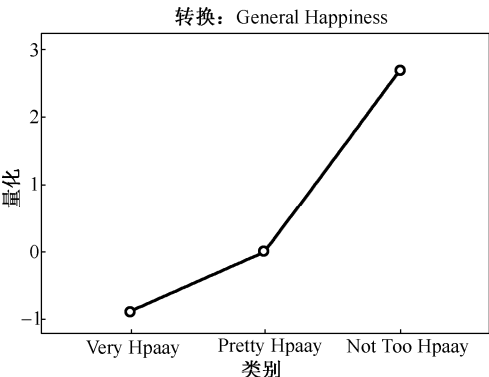


图 11-88 happy 转换图

表 11-85 所示为相关系数与容忍度统计量表，它列出了零阶相关、偏相关和部分相关系数。从重要性指标可见，回归方程中对因变量最重要的自变量为 life，其余 3 个自变量对因变量不太重要。从部分相关系数中也可看到，当 life 进入回归方程后，复相关系数的平方增加为 0.370，也说明该变量对因变量比较重要。从容忍度列中可见，各变量的容忍度都大于 0.1，因此，变量之间不存在多重共线性。

从图 11-88 可见，用有序样条变换后，happy 3 个类别的原先顺序得以保留，但类别值已发生变化(原分别为 1，2，3)。

11.11 对数线性模型

11.11.1 对数线性模型的概念

1. 概述

对数线性模型适用于分析列联表数据。与描述概率  $p$  与协变量  $x_1, \dots, x_p$  之间关系的 Logistic 模型

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

相比，对数线性模型

$$\ln m = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

描述了期望频数  $m$  与协变量  $x_1, \dots, x_p$  之间的关系。它们都是广义线性模型的一种特殊形式。

对数线性模型假定列联表中的单元格频数服从泊松分布或多项式分布，可以引用于任意维数的列联表。

在对数线性模型应用于列联表数据的分析中，每个类别是一个响应变量。与回归分析相比，对数线性分析更像相关分析。其核心是对每对变量间的关联关系进行研究，而非在其他项上对它们中的一个响应类别构建模型。

为更清楚地理解对数线性模型在多维列联表中是如何应用的，不妨以三维列联表为例来加以说明，因为三维以上列联表的对数线性回归分析与其极其相似。

假设有  $n$  个被试对象是根据属性  $A$ 、 $B$ 、 $C$  来进行分类的，它们依次有  $I$  个水平、 $J$  个水平、 $K$  个水平， $n$  个被试对象中属于  $A_i$ 、 $B_j$ 、 $C_k$  类的有  $n_{ijk}$  个，则可得到如表 11-86 所示的三维  $I \times J \times K$  列联表。

表 11-86 三维  $I \times J \times K$  列联表

C	A	B		
		$B_1$	...	$B_J$
$C_1$	$A_1$	$n_{111}$	...	$n_{1J1}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$A_I$	$n_{I11}$	...	$n_{IJ1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C_K$	$A_1$	$n_{11K}$	...	$n_{1JK}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$A_I$	$n_{I1K}$	...	$n_{IJK}$

对数线性模型描述的是列联表分表中的条件关联关系，也就是在控制其中一个变量(称第 3 个变量)时，其他两个变量之间的关联关系。当每个分表的总体满足独立性时，称每对变量是条件独立的，此时，分表中的优势比为 1。下面按关联程度由低到高的顺序，给出关于列联表中 5 个层次的对数线性模型。

(1) 3 对变量全部是条件独立，即控制  $C$  则  $A$  与  $B$  是独立的；控制  $B$  则  $A$  与  $C$  是独立的；控制  $A$  则  $B$  与  $C$  是独立的。

- (2) 3 对变量中有两对是条件独立。例如，控制  $C$  则  $A$  与  $B$  是独立的；控制  $B$  则  $A$  与  $C$  是独立的；控制  $A$  则  $B$  与  $C$  是关联的。
- (3) 3 对变量中有一对是条件独立。例如，控制  $C$  则  $A$  与  $B$  是独立的；控制  $B$  则  $A$  与  $C$  是关联的；控制  $A$  则  $B$  与  $C$  是关联的。
- (4) 没有一对变量是条件独立，但在第 3 个变量的每个类别上，任意两个变量之间的关联关系都是一样的，称其为同质性关联。
- (5) 每对变量都是关联的，且有交互效应，即每对变量之间的关联关系由于第 3 个变量的类别不同而不同。

下面用符号来表示模型，符号中有连在一起的，表明这些变量之间存在关联关系。例如，用符号  $(A, B, C)$  表示上面(1)中的模型， $A$ 、 $B$ 、 $C$  没有连在一起，表明这 3 个变量是条件独立的。再如， $(AB, AC, BC)$  用来表示上面(4)中的模型，表示所有 3 对变量都是关联的。符号  $(ABC)$  表示(5)中的模型，由于这个模型能准确无误地拟合三维列联表的样本数据，故称这个模型为饱和模型。而其他模型只是包含饱和模型的参数子集的更简单模型，故称为简约模型。

三维列联表中各种情况下的典型对数线性模型列在表 11-87 中。

表 11-87 三维列联表中的各种情况下的典型对数线性模型

模型符号	对数线性模型	说 明
$(A, B, C)$	$\ln m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$	完全独立
$(AB, C)$	$\ln m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB}$	部分独立
$(AB, BC)$	$\ln m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC}$	条件独立
$(AB, BC, AC)$	$\ln m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC}$	同质性关联
$(ABC)$	$\ln m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC}$	饱和模型

表中的  $m_{ijk}$  为期望频数， $m_{ijk} = E(n_{ijk})$ ， $i = 1, \cdots, I$ ； $j = 1, \cdots, J$ ； $k = 1, \cdots, K$ ； $\lambda_i^A$ 、 $\lambda_j^B$ 、 $\lambda_k^C$  分别表示变量  $A$ 、 $B$ 、 $C$  的主效应； $\lambda_{ij}^{AB}$ 、 $\lambda_{jk}^{BC}$ 、 $\lambda_{ik}^{AC}$  分别表示  $A$ 、 $B$ 、 $C$  两两间的交互效应，在 SPSS 的对数线性模型中，将其称为二阶交互效应；而  $\lambda_{ijk}^{ABC}$  表示的是  $A$ 、 $B$ 、 $C$  三者的交互效应，SPSS 的模型中将其称为三阶交互效应。表中模型等式右边的各被加项称为参数。

通常采用最大似然估计法来对对数模型中的参数进行估计。理论上可以证明，在单元格频数服从泊松分布或多项分布时，它们的对数线性模型的参数具有相同的最大似然估计。由于模型参数较多，因此要想得到模型参数唯一的最大似然估计，需对参数施加条件约束。在 SPSS 中，对于每个变量的最后一个分类的参数被设置为 0，模型中的交互作用项等同于一个新变量。称被设置为 0 的参数是冗余的。

用对数线性模型来分析变量之间有无关联关系时，实质上就是在检验交互作用项对应参数是否等于 0，若不能拒绝这样的假设，则认为变量间相互独立，否则变量间存在关联关系。关联关系的强弱可用优势比来加以描述，优势比大于 1 表明变量间存在正关联，等于 1 表明变量间独立，小于 1 表明变量间存在负关联。

对模型拟合效果的检验，采用 Pearson 卡方检验和似然比检验。Pearson 卡方检验统计量的计算公式为

$$\chi^2 = \sum \frac{(\text{观测频数} - \text{期望频数})^2}{\text{期望频数}}$$

似然比检验统计量为

$$G^2 = 2 \sum \text{观测频数} \times \lg \frac{\text{观测频数}}{\text{期望频数}}$$

如果统计量的  $p$  值大于 0.05, 则模型拟合较好, 否则模型拟合不佳。

## 2. SPSS 的对数线性模型过程对数据的要求

(1) 因变量只能是分类变量, 最多可以选择 10 个因变量。  
(2) 因子只能是分类变量, 最多可以选择 10 个因子来定义列联表的单元格。单元格中的观测值称为单元格频数。

(3) 单元协变量为连续型变量。

(4) 单元结构变量用来指定变量的权重。当分类变量的某些组合不可能存在时, 则对应的单元格是单元结构中的无效单元格, 其单元结构变量值为 0 或 1。不可以使用单元结构变量对分类汇总数据进行加权, 而应从【数据】菜单中选择【加权个案】来进行加权。

当列联表中有结构 0 存在时, 将此类列联表称为不完全列联表。SPSS 用【单元结构】选项来对不完全列联表进行识别, 在【单元结构】中的变量值为非正数时, 认为是结构 0 数据。在样本量不大时, 表格数较多的表的单元格中也会出现观测值 0, 这称为抽样 0, 因此, SPSS 将列联表中的单元格频数 0 默认为抽样 0; 否则, 需在单元结构加权时, 对 0 作出定义。

(5) 对比变量为连续型变量。它们用来计算广义对数几率的比值。对比变量的值是期望单元格频数的对数线性组合的系数。

## 3. 对数线性模型的过程

在 SPSS 中, 单击【分析】菜单中的【对数线性模型】(见图 11-89), 它提供了 3 个用来进行对数线性模型分析的过程: 【常规】(一般对数线性回归)过程、【Logit】(Logit 对数线性回归)过程与【模型选择】过程。它们分别适用于不同的研究场合。虽然它们的算法略有不同, 但参数估计的结果是一样的, 用来对参数进行估计的方法均为 Newton-Raphson 法。

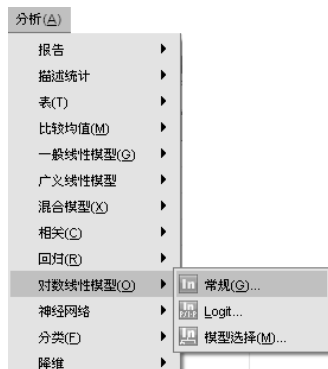


图 11-89 对数线性模型过程

### 11.11.2 一般对数线性回归分析

在建立分层或非分层的对数线性模型中, 均可用一般对数线性回归过程。本过程是一个证实性研究过程, 研究人员使用本过程时, 应对数据有较多的了解, 已经知道需要建立什么样的模型, 对应检验的参数是什么等信息, 拟合模型的目的是为了验证原先经验结论的正确性。

在一般对数线性回归过程中, 没有因变量和自变量之分, 进入模型的分变量都作为影响单元格频数的因子(因素)对待。

SPSS 在一般对数线性分析中提供两种分布: 泊松分布和多项分布。

在泊松分布假设下, 研究前不需确定总的样本量, 或分析不依赖于总的样本量。单元格的观测频数之间相互独立。

在多项分布假设下, 总的样本量是固定的, 或分析依赖于总的样本量。单元格的观测频数之间相互不独立。

对于对数线性模型, 一般而言, 真正有用的模型不是饱和模型, 而是不饱和的简约模型。

但饱和模型只有一个，而简约模型却有多个，而且随着列联表维数以及各变量水平数的增多而成倍增加。如何才能找到简约模型呢？显然不同的研究人员有不同的作法，不妨也可以这样来处理：

先建立饱和模型和，通过检验每个参数的  $Z$  值或置信区间，从饱和模型中剔除无意义的效应；再建立主效应模型，通过查看拟合优度检验结果，来判定所建模型是否有意义；从饱和模型出发用淘汰法，从主效应模型出发用加入法，逐渐寻找最佳简约模型。在各个参数均有统计学显著性意义的前提下，拟合优度值最小的简约模型较佳。

1. 一般线性对数回归过程

- (1) 按【分析→对数线性模型→常规】顺序，打开【常规对数线性分析】对话框，见图 11-90。
- (2) 选定模型中需要的各种变量。  
从左侧的源变量框中选择多个分类变量作为因子变量进入【因子】框中。  
从左侧的源变量框中选择一个或多个连续型变量作为单元协变量进入【单元协变量】框中。  
从左侧的源变量框中选择一个单元结构变量进入【单元结构】框中。用来定义单元格中是否含有结构 0 的单元格。  
从左侧的源变量框中选择一个或多个连续型的对比变量进入【对比变量】框中。
- (3) 选择单元格频数的分布类型。在【单元计数分布】栏中，根据研究问题的实际情况，选择单元格频数的分布是【泊松】分布还是【多项式分布】。
  - ①【泊松】分布。如果单元格的观测频数之间相互独立，选择本项。
  - ②【多项式分布】。如果单元格的观测频数之间相互不独立，选择本项。
- (4) 设定模型。单击【模型】按钮，打开如图 11-91 所示的【常规对数线性分析：模型】对话框。

- ①【指定模型】栏。指定模型类型。
  - 【饱和】模型。在模型中包含在【因子】框中所选因子变量的所有主效应和交互效应。在饱和模型中，不包含协变量项。它是系统默认选项。



图 11-90 【常规对数线性分析】对话框



图 11-91 【常规对数线性分析：模型】对话框

- 【设定】(定制)。用户需自己定义模型中所要用到的交互项，包括因子与协变量之间的交互项。
- ②【因子与协变量】框。列出【常规对数线性分析】对话框中所选取的因子和协变量的名称。
- ③ 在用户选择【设定】之后，【模型中的项】框被激活，选定模型所需变量或变量组合，单击【构建项】栏中的下拉列表，选择变量在模型中所取的作用，参见 11.5.2 节第(7)项【设



定】中的相关内容。用户从因子与协变量框中所选择的进入自制模型的变量名称(主效应)及变量组合名称(交互效应)指定后,所选项将显示在本框中。值得一提的是:必须指定在模型中需要包含的所有项。

单击【继续】按钮,返回【常规对数线性分析】对话框。

(5) 单击【选项】按钮,打开如图 11-92 所示的【常规对数线性分析:选项】对话框,可以选择输出有关模型信息、拟合优度、单元格期望频数、残差等统计量、统计图,还可以选择拟合过程中的迭代收敛的标准。

#### ① 【输出】栏。

- 【频率】(应为频数)。输出频数表。此项为系统默认选项。
- 【残差】。在输出表中包含残差项信息。此项为系统默认选项。
- 【设计矩阵】。输出设计矩阵表。
- 【估计】。输出模型的参数估计表。此项为系统默认选项。
- 【迭代历史记录】。输出模型的迭代历史记录表。

#### ② 【图】栏。

- 【调节的残差值】(应为调整残差图)。输出调整残差图。
- 【调节残差值的正态概率】(应为调整残差的正态概率图)。输出调整残差的正态概率图。
- 【偏差残差】。输出 Deviance 残差图。

Deviance 残差的计算公式为

$$d_i = \text{sgn}(O_i - E_i) \left[ 2O_i \lg \frac{O_i}{E_i} + 2(n_i - O_i) \lg \frac{n_i - O_i}{n_i - E_i} \right]^{\frac{1}{2}}$$

式中,  $O_i$  为观测频数;  $E_i$  为期望频数;  $n_i$  为每一自变量组合的观测单位数。

- 【偏差的正态概率】。输出 Deviance 残差的正态概率图。

③ 【置信区间】框。输入一个 0~100 之间的值,可以调整参数估计值的置信区间。系统默认值为 95。

#### ④ 【标准】栏。使用 Newton-Raphson 方法来获取最大似然参数估计值。

- 【最大迭代】次数。系统默认值为 20。
- 【收敛性】标准。可在其下拉列表中进行选择。系统默认值为 0.001。
- 【Delta】。设置饱和模型的校正系数。系统默认值为 0.5。

单击【继续】按钮,返回【常规对数线性分析】对话框。

(6) 单击【保存】按钮,打开如图 11-93 所示的【常规对数线性分析:保存】对话框,选取要在活动数据集中保存为新变量的各种值。新变量名称中的后缀  $n$  会递增,以使每个保存变量都具有唯一的名称。

可以保存 4 种类型的残差:原始残差、标准化残差、调整残差和 Deviance 残差,分别对应【残差】、【标准残差值】、【调节的残差值】、【偏差残差】;还可以保存【预测值】。



图 11-92 【常规对数线性分析:选项】对话框

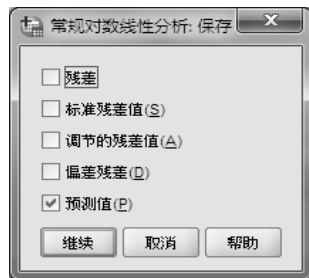


图 11-93 【常规对数线性分析:保存】对话框

原始残差 = 观测频数 - 期望频数，所以它是观测值与期望值之差。  
标准化残差为

$$\frac{\text{残差}}{\sqrt{\text{期望频数} \times \left(1 - \frac{\text{期望频数}}{n}\right)}}$$

式中， $n$  为样本量。  
由于调整残差的计算比较复杂，在此不列出其计算公式，有兴趣的读者可参阅一般对数线性回归中残差算法方面的资料。

2. 一般对数线性回归实例分析

【例 14】 美国赖特州立大学医学院以及俄亥俄州代顿的联合健康服务机构，于 1992 年对来自代顿附近郊区的 2276 名高中学生进行了有关是否有饮酒、吸烟或使用大麻的情况调查。调查结果见表 11-87，对应数据文件为 data11-12，并使用其中的频数变量做过加权处理。

现用一般对数线性模型分析该地区高中学生喝酒、吸烟和使用大麻 3 种行为是否存在关联关系。

1) 建立饱和模型

- (1) 打开数据文件 data11-12。按【分析→对数线性模型→常规】顺序，打开【常规对数线性分析】对话框。
- (2) 从左侧的源变量框中选择饮酒、吸烟、使用大麻 3 个分类变量作为因子变量进入【因子】框。
- (3) 其他选用系统默认选项，即作饱和模型分析。
- (4) 单击【确定】按钮，则在输出窗中得到相关的模型拟合信息。只需先看其中的拟合度检验表(见表 11-88)、单元格频数和残差表(见表 11-89)和参数估计表(见表 11-90)。

表 11-87 高中学生饮酒、吸烟或使用大麻的调查结果

饮酒	吸烟	使用大麻	
		是	否
是	是	911	538
	否	44	456
否	是	3	43
	否	2	279

表 11-88 拟合优度检验表

a,b			
	值	df	Sig.
似然比	.000	0	.
Pearson 卡方检验	.000	0	.

模型：泊松  
设计:常量 + 饮酒 + 吸烟 + 使用大麻 + 饮酒 \* 吸烟 + 饮酒 \* 使用大麻 + 吸烟 \* 使用大麻 + 饮酒 \* 吸烟 \* 使用大麻

表 11-89 观测频数与期望频数表

单元计数和残差 <sup>a,b</sup>										
饮酒	吸烟	使用大麻	观测		期望的		残差	标准化残差	调整残差	偏差
			计数	%	计数	%				
否	否	否	279.500	12.3%	279.500	12.3%	.000	.000	.	.000
		是	2.500	0.1%	2.500	0.1%	.000	.000	.000	.000
	是	否	43.500	1.9%	43.500	1.9%	.000	.000	.000	.000
		是	3.500	0.2%	3.500	0.2%	.000	.000	.000	.000
是	否	否	456.500	20.0%	456.500	20.0%	.000	.000	.	.000
		是	44.500	2.0%	44.500	2.0%	.000	.000	.000	.000
	是	否	538.500	23.6%	538.500	23.6%	.000	.000	.000	.000
		是	911.500	40.0%	911.500	40.0%	.000	.000	.	.000

a. 模型：泊松  
b. 设计:常量 + 饮酒 + 吸烟 + 使用大麻 + 饮酒 \* 吸烟 + 饮酒 \* 使用大麻 + 吸烟 \* 使用大麻 + 饮酒 \* 吸烟 \* 使用大麻

表 11-90 参数估计表

参数估计 <sup>b,c</sup>						
参数	估计	标准误	Z	Sig.	95% 置信区间	
					下限	上限
常量	6.815	.033	205.755	.000	6.750	6.880
[饮酒 = 0]	-5.562	.536	-10.386	.000	-6.612	-4.513
[饮酒 = 1]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 0]	-3.020	.154	-19.669	.000	-3.320	-2.719
[吸烟 = 1]	0 <sup>a</sup>	.	.	.	.	.
[使用大麻 = 0]	-.526	.054	-9.683	.000	-.633	-.420
[使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 0] * [吸烟 = 0]	2.683	.842	3.186	.001	1.032	4.334
[饮酒 = 0] * [吸烟 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 0]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 0] * [使用大麻 = 0]	3.046	.558	5.457	.000	1.952	4.140
[饮酒 = 0] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [使用大麻 = 0]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 0] * [使用大麻 = 0]	2.854	.166	17.176	.000	2.529	3.180
[吸烟 = 0] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 1] * [使用大麻 = 0]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 1] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 0] * [吸烟 = 0] * [使用大麻 = 0]	-.658	.860	-.765	.445	-2.344	1.028
[饮酒 = 0] * [吸烟 = 0] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 0] * [吸烟 = 1] * [使用大麻 = 0]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 0] * [吸烟 = 1] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 0] * [使用大麻 = 0]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 0] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 1] * [使用大麻 = 0]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 1] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.

a. 此参数为冗余参数，因此将被设为零。  
b. 模型：泊松  
c. 设计:常量 + 饮酒 + 吸烟 + 使用大麻 + 饮酒 \* 吸烟 + 饮酒 \* 使用大麻 + 吸烟 \* 使用大麻 + 饮酒 \* 吸烟 \* 使用大麻

(5) 结果分析。

从表 11-88 可见，在饱和模型下，似然比值为 0.000。从表 11-89 可见，对应各单元格中的观测频数与期望频数全都相等，表明模型能全部正确预测观测值。

但从表 11-90 中可见，饮酒、吸烟、使用大麻的交互作用项的 Sig.=0.445，大于 0.05，现有证据不足以拒绝三阶交互作用项的参数为 0 的假设，故不能认为饮酒、吸烟、使用大麻之间两两都是关联的。该三阶交互作用项可以从模型中移去。

2) 建立主效应模型

(1)、(2)步操作方法同上。

(3) 单击【模型】按钮，打开【常规对数线性分析：模型】对话框。在【指定模型】栏中，选择【设定】选项。在【因子与协变量】框中一次性选定饮酒、吸烟、使用大麻变量，单击【构建项】下拉列表，选择【主效应】，单击右移箭头，将饮酒、吸烟、使用大麻变量移入【模型中的项】框中。单击【继续】按钮，返回【常规对数线性分析：模型】对话框。

(4) 其他选用系统默认选项。

(5) 单击【确定】按钮，则在输出窗中得到相关的模型拟合信息。在此，只需先看其中的拟合度检验表(见表 11-91)的结果。

(6) 结果分析。

从表 11-91 可见，在主效应模型下，似然比值为

表 11-91 拟合优度检验表<sup>a, b</sup>

	值	df	Sig.
似然比	1286.020	4	.000
Pearson 卡方检验	1411.386	4	.000

模型：泊松  
设计：常量 + 饮酒 + 吸烟 + 使用大麻

1286.020, Sig.=0.000, 小于 0.05, 因此, 主效应模型拟合效果不佳。这表明, 当从饱和模型中移去所有二阶交互作用项时, 模型已发生显著变化。因此, 应在模型中考虑二阶交互效应。

3) 建立含二阶交互效应的模型

(1)、(2)步操作方法同上。

(3) 单击【模型】按钮, 打开【常规线性分析: 模型】对话框。在【指定模型】栏中, 选择【设定】选项。在【因子与协变量】框中一次性选定饮酒、吸烟、使用大麻变量, 单击【构建项】的下拉列表, 选择【主效应】, 单击右移箭头, 将饮酒、吸烟、使用大麻变量移入【模型中的项】框中。再在【因子与协变量】框中一次性选定饮酒、吸烟、使用大麻变量, 单击【构建项】的下拉列表, 选择【所有二阶】选项, 单击右移箭头, 将 3 个变量两两间的交互效应都选入模型中。单击【继续】按钮, 返回【常规线性分析: 模型】对话框。

(4) 单击【选项】按钮, 打开【常规线性分析: 选项】对话框。在【输出】选项中除保留默认选项外, 再选择【估计】选项。单击【继续】按钮, 返回【常规线性分析: 模型】对话框。

(5) 其他选用系统默认选项。

(6) 单击【确定】按钮, 则在输出窗中得到相关的模型拟合信息。

(7) 结果分析。

表 11-92 数据基本信息

数据信息		N
案例	有效	8
	缺失	0
	加权有效	2276
单元格	定义的单元格	8
	结构中的无效单元	0
	采样无效单元	0
类别	饮酒	2
	使用大麻	2

表 11-93 迭代信息

收敛信息 <sup>a,b</sup>	
最大迭代次数	20
收敛容限度	.00100
最终最大绝对差值	.00199
最终最大相对差值	.00067 <sup>c</sup>
迭代次数	8
a. 模型: 泊松	
b. 设计:常量 + 饮酒 + 吸烟 + 使用大麻 + 吸烟 * 使用大麻 + 饮酒 * 使用大麻 + 饮酒 * 吸烟	
c. 由于参数估计的最大相对变化小于指定的收敛条件, 导致迭代已收敛。	

表 11-92 所示是数据的基本信息。样本量为 2276, 单元格为 8 个, 没有结构 0 或抽样 0 数据出现。

表 11-94 拟合优度检验表

a,b			
	值	df	Sig.
似然比	.374	1	.541
Pearson 卡方检验	.401	1	.527

模型: 泊松  
设计:常量 + 饮酒 + aa + 使用大麻 + aa \* 使用大麻 + 饮酒 \* aa + 饮酒 \* 使用大麻

表 11-93 所示是参数估计拟合过程中的迭代信息。设定最大迭代次数 20 次, 实际迭代 8 次达到收敛容忍度标准 0.001, 最终最大绝对差值为 0.00199, 最终最大相对差值为 0.00067。

表 11-94 所示是拟合优度检验结果。似然比检验和 Pearson 卡方检验的 Sig.值都大于 0.05, 表明模型对数据拟合得较好。与只有主效应的模型相比, 当模型中引入 3 个

变量的两两交互效应后, 模型的拟合效果有显著变化。

表 11-95 所示为观测频数、期望频数、残差等信息。由于残差值不大, 也进一步说明本模型的拟合效果不错。

表 11-96 所示为模型中各参数的估计、标准误、Z 值、P 值和 95%的置信区间等信息。各

参数估计的  $Z$  检验  $p$  值均小于 0.01, 有足够证据拒绝这些参数为 0 的假设, 表明模型中加入这些参数对应的变量是有效的。

表 11-95 观测频数与期望频数表

			单元计数和残差 <sup>a,b</sup>							
饮酒	吸烟	使用大麻	观测		期望的		残差	标准化残差	调整残差	偏差
			计数	%	计数	%				
否	否	否	279	12.3%	279.617	12.3%	-.617	-.037	-.633	-.037
		是	2	0.1%	1.383	0.1%	.617	.524	.633	.491
	是	否	43	1.9%	42.383	1.9%	.617	.095	.633	.095
		是	3	0.1%	3.617	0.2%	-.617	-.324	-.632	-.334
是	否	否	456	20.0%	455.383	20.0%	.617	.029	.633	.029
		是	44	1.9%	44.617	2.0%	-.617	-.092	-.633	-.093
	是	否	538	23.6%	538.617	23.7%	-.617	-.027	-.633	-.027
		是	911	40.0%	910.383	40.0%	.617	.020	.633	.020

a. 模型: 泊松

b. 设计: 常量 + 饮酒 + 吸烟 + 使用大麻 + 吸烟 \* 使用大麻 + 饮酒 \* 使用大麻 + 饮酒 \* 吸烟

表 11-96 参数估计表

参数估计 <sup>b,c</sup>						
参数	估计	标准误	Z	Sig.	95% 置信区间	
					下限	上限
常量	6.814	.033	205.699	.000	6.749	6.879
[饮酒 = 0]	-5.528	.452	-12.237	.000	-6.414	-4.643
[饮酒 = 1]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 0]	-3.016	.152	-19.891	.000	-3.313	-2.719
[吸烟 = 1]	0 <sup>a</sup>	.	.	.	.	.
[使用大麻 = 0]	-.525	.054	-9.669	.000	-.631	-.418
[使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 0] * [使用大麻 = 0]	2.848	.164	17.383	.000	2.527	3.169
[吸烟 = 0] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 1] * [使用大麻 = 0]	0 <sup>a</sup>	.	.	.	.	.
[吸烟 = 1] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 0] * [使用大麻 = 0]	2.986	.464	6.432	.000	2.076	3.896
[饮酒 = 0] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [使用大麻 = 0]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [使用大麻 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 0] * [吸烟 = 0]	2.055	.174	11.804	.000	1.713	2.396
[饮酒 = 0] * [吸烟 = 1]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 0]	0 <sup>a</sup>	.	.	.	.	.
[饮酒 = 1] * [吸烟 = 1]	0 <sup>a</sup>	.	.	.	.	.

a. 此参数为冗余参数, 因此将被设为零。

b. 模型: 泊松

c. 设计: 常量 + 饮酒 + 吸烟 + 使用大麻 + 吸烟 \* 使用大麻 + 饮酒 \* 使用大麻 + 饮酒 \* 吸烟

结合表 11-94 中得到的结论, 可以认为 3 个变量的两两交互效应确实存在。

表 11-95 中的期望频数是用表 11-96 中的估计计算得到的。下面以计算  $n_{101}$  的观测频数的期望频数  $m_{101} = E(n_{101})$  为例来说明这两表之间的关系。

从数据文件 data11-12 中可知, 饮酒、吸烟、使用大麻变量的值标签中, “1” 表示是, “0” 表示否, 因此,  $n_{101}$  是指表中饮酒为 “是”、吸烟为 “否”、使用大麻为 “是” 所对应单元格中的观测频数, 其值为 44。有

$$\begin{aligned} \ln m_{101} &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = \lambda + \lambda_1^A + \lambda_0^B + \lambda_1^C + \lambda_{10}^{AB} + \lambda_{01}^{BC} + \lambda_{11}^{AC} \\ &= 6.814 + 0 - 3.016 + 0 + 0 + 0 + 0 = 3.798 \end{aligned}$$

所以 
$$m_{101} = e^{3.798} = 44.617$$

这与表 11-95 中的期望频数 44.617 是一致的。

基于表 11-96 中的交互作用项的参数估计，依据交互项的参数估计可用来描述条件关联，还可得到估计的优势比。由于本模型假定的是同质性关联，因此，估计的优势比在给定变量的每个类别上是相同的。

由此可得，在饮酒的每个水平上，吸烟和使用大麻的优势比为  $e^{2.848} = 17.3$ ；在吸烟的每个水平上，饮酒和使用大麻的优势比为  $e^{2.986} = 19.8$ ；在使用大麻的每个水平上，吸烟和饮酒的优势比为  $e^{2.055} = 7.8$ 。由于各优势比的值都远大于 1，说明每对变量间估计的条件正关联关系是非常强的。

事实上，在不同的模型下，可得到各自模型下的条件优势比，而且所得到的条件优势比往往是不一样的。因此，选择一个好的模型是最重要的。

重复上述过程，把处在主效应模型与饱和模型之间的所有可能的模型都运行一遍，利用拟合优度检验结果，则可寻得所有模型中的最优模型。

读者可自行加以验证，在本例中， $(AB, BC, AC)$  模型就是最佳简约模型。

表 11-97 所示为参数估计的相关系数矩阵。

表 11-97 参数估计的相关系数表

参数估计的相关性 <sup>a,b,c</sup>							
	常量	[饮酒 = 0]	[吸烟 = 0]	[使用大麻 = 0]	[吸烟 = 0] * [使用大麻 = 0]	[饮酒 = 0] * [使用大麻 = 0]	[饮酒 = 0] * [吸烟 = 0]
常量	1	-.054	-.214	-.609	.197	.051	.006
[饮酒 = 0]	-.054	1	-.074	.013	.083	-.941	-.105
[吸烟 = 0]	-.214	-.074	1	.125	-.922	.080	-.026
[使用大麻 = 0]	-.609	.013	.125	1	-.324	-.084	.187
[吸烟 = 0] * [使用大麻 = 0]	.197	.083	-.922	-.324	1	-.065	-.113
[饮酒 = 0] * [使用大麻 = 0]	.051	-.941	.080	-.084	-.065	1	-.204
[饮酒 = 0] * [吸烟 = 0]	.006	-.105	-.026	.187	-.113	-.204	1

- a. 模型：泊松
- b. 设计:常量 + 饮酒 + 吸烟 + 使用大麻 + 吸烟 \* 使用大麻 + 饮酒 \* 使用大麻 + 饮酒 \* 吸烟
- c. 未显示冗余的参数。

表 11-98 参数估计的协方差系数表

参数估计的协方差 <sup>a,b,c</sup>							
	常量	[饮酒 = 0]	[吸烟 = 0]	[使用大麻 = 0]	[吸烟 = 0] * [使用大麻 = 0]	[饮酒 = 0] * [使用大麻 = 0]	[饮酒 = 0] * [吸烟 = 0]
常量	.001	-.001	-.001	-.001	.001	.001	.000
[饮酒 = 0]	-.001	.204	-.005	.000	.006	-.197	-.008
[吸烟 = 0]	-.001	-.005	.023	.001	-.023	.006	-.001
[使用大麻 = 0]	-.001	.000	.001	.003	-.003	-.002	.002
[吸烟 = 0] * [使用大麻 = 0]	.001	.006	-.023	-.003	.027	-.005	-.003
[饮酒 = 0] * [使用大麻 = 0]	.001	-.197	.006	-.002	-.005	.216	-.016
[饮酒 = 0] * [吸烟 = 0]	.000	-.008	-.001	.002	-.003	-.016	.030

- a. 模型：泊松
- b. 设计:常量 + 饮酒 + 吸烟 + 使用大麻 + 吸烟 \* 使用大麻 + 饮酒 \* 使用大麻 + 饮酒 \* 吸烟
- c. 未显示冗余的参数。

表 11-98 所示为参数估计的协方差矩阵。

在输出窗中，另外还有 3 张诊断图。图 11-94 所示为 8 个单元格的观测频数、期望频数和

校正残差的散点图。由散点图可见，8 个散点明显存在一定的趋势。图 11-95、图 11-96 所示的 Q-Q 图中，散点很有规则地分布在直线的两侧，说明残差不服从正态分布。

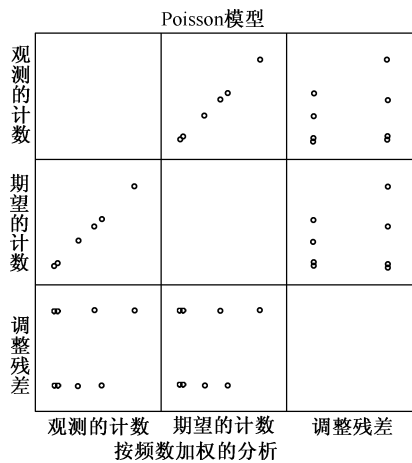


图 11-94 观测频数、期望频数和调整残差两两对应的散点图

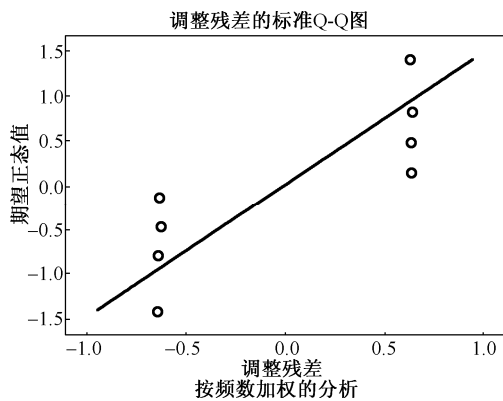


图 11-95 校正残差的正态 Q-Q 图

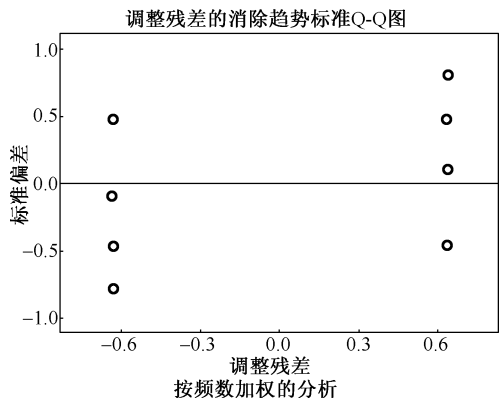


图 11-96 校正残差的去势正态 Q-Q 图

11.11.3 Logit 对数线性回归分析

一般对数线性模型中，对列联表分类变量之间的因果关系不需要了解，关注的是分类变量之间的条件关联关系。如果分类变量之间的因果关系明确，就需要研究因变量(或响应变量)与自变量(或解释变量)之间的关系。此时，就需要使用 Logit 对数线性回归分析。它所用的参数估计方法也是 Newton-Raphson 法。

名义响应变量的 Logit 模型将每个类别与一个基准类别配比。SPSS 软件中使用类别值最大的那个类别作为基准类别。

为便于理解，假设作响应变量的分类变量共有  $k$  个类别，在最简单只有一个预测变量  $x$  的基准类别的 Logit 模型为

$$\ln \frac{p(y = j)}{p(y = k)} = \alpha_j + \beta_j x \quad j = 1, \dots, k - 1$$

在已知响应类别落在  $j$  或最后一个类别中，它构建的是响应类别为  $j$  的几率的对数。Logit 本身就是由 Logit 组成的，其本意就是对它取对数。

在 Logit 模型中，自动采用多项分布，并假定多项分布中的观测值是独立的。  
在 SPSS 中参考水平对应的参数被设置为冗余参数，模型拟合时，系统自动将其置为 0，且不对其进行检验。

Logit 对数线性回归过程除了分析人为指定引入模型的各项外，还将自动对引入的这些项与因变量的交互项进行分析，而不管用户在模型选项中是如何设定的。关于这一点，在程序运行后的输出结果中将会感受到。

值得一提的是，在拟合结果上，Logit 模型与前面介绍过的 Logistic 模型等价。

1. Logit 对数线性回归过程

按【分析→对数线性模型→Logit】顺序，打开【Logit 对数线性分析】对话框，见图 11-97。



图 11-97 【Logit 对数线性分析】对话框

在该对话框中，除增加了【因变量】框，以及比【常规对数线性分析】对话框中少了一个需要选择【单元计数分布】选项外，其他与【常规对数线性分析】对话框没有区别。

单击【保存】、【模型】、【选项】按钮打开的各对话框的界面、选项等与【常规对数线性分析】对话框也都一样。

因此，为节省篇幅，重复部分请参见第 11.11.2 节相关内容。

值得一提的是，在【Logit 对数线性分析】对话框中，需选取一个可以有多个类别的分类变量作为分析中的因变量，将其移入【因变量】框中。

其他变量作为自变量分别移到其相应的框中。这里的因子变量也必须为分类变量，最多可输入 10 个，在这里也是作为自变量来使用的。

2. Logit 对数线性回归实例分析

【例 15】 一项在不同性别和种族的人群中开展的“相信死后有来世”的社会调查，得到了表 11-99 所示的调查结果。设  $y$  = 相信死后有来世，用性别、种族作自变量，对因变量  $y$  进行 Logit 对数线性回归分析。

表 11-99 性别、种族与相信来世

种族	性别	相信死后有来世		
		是	不确定	否
黑人	女	64	9	15
	男	25	5	13
白人	女	371	49	74
	男	250	45	71

表 11-99 对应的数据文件为 data11-13。“种族”（值标签：0—黑人，1—白人）、“性别”（值标签：0—女，1—男）、“相信死后有来世”（值标签：1—是，2—不确定，3—否）均为名义分类变量。



操作步骤如下：

- (1) 打开数据文件 data11-13。按【分析→对数线性模型→Logit】顺序，打开【Logit 对数线性分析】对话框。
- (2) 从左侧的源变量框中选择“相信死后有来世”变量作为因变量进入【因变量】框。
- (3) 从左侧的源变量框中选择“种族”、“性别”两个分类变量作为因子变量进入【因子】框。
- (4) 单击【Logit 对数线性分析：模型】按钮，打开模型对话框。在【指定模型】栏中，选择【设定】选项。在【因子与协变量】框中一次性选定“种族”、“性别”变量，单击【构建项】栏的下拉列表，选择【主效应】，单击右移箭头，将“种族”、“性别”变量移入【模型中的项】框。单击【继续】按钮，返回【Logit 对数线性分析】对话框。
- 注意：选择主效应选项的原因是由于性别和种族变量之间不存在交互效应。
- (5) 单击【选项】按钮，在打开的对话框的【输出】栏中，选择【估计】选项。其他选用系统默认选项。单击【继续】按钮，返回【Logit 对数线性分析】对话框。
- (6) 单击【确定】按钮，在输出窗中得到相关的模型拟合信息。
- (7) 结果分析。

表 11-100 所示为参与到模型中的数据基本情况，表 11-101 所示为拟合模型过程中的迭代信息。对这两表的解释参见【例 14】的结果分析。

表 11-102 所示为拟合优度检验结果，由于似然比和 Pearson 卡方检验统计量值都不足 1，且显著性 *p* 值都远大于 0.05，因此，可以认为模型对原始数据拟合良好。

表 11-103 所示为对因变量的离散性分析。在该表中出现了两个新的统计量，一个是熵，另一个是集中度。这两个统计量在分类资料中都是用来描述数据资料的离散程度的。

表 11-100 数据基本信息

数据信息		N
案例	有效	12
	缺失	0
	加权有效	991
单元格	定义的单元格	12
	结构中的无效单元	0
	采样无效单元	0
类别	相信死后有来世	3
	种族	2

表 11-101 迭代信息

收敛信息 <sup>a,b</sup>	
最大迭代次数	20
收敛容限度	.00100
最终最大绝对差值	.00008 <sup>c</sup>
最终最大相对差值	.00015
迭代次数	4
a. 模型：多项 Logit	
b. 设计：常量 + 相信死后有来世 + 相信死后有来世 * 种族 + 相信死后有来世 * 性别	
c. 由于参数估计的最大绝对变化小于指定的收敛条件，导致迭代已收敛。	

表 11-102 拟合优度检验

拟合度检验 <sup>a,b</sup>			
	值	df	Sig.
似然比	.854	2	.653
Pearson 卡方检验	.861	2	.650
a. 模型：多项 Logit			
b. 设计：常量 + 相信死后有来世 + 相信死后有来世 * 种族 + 相信死后有来世 * 性别			

表 11-103 离散性分析

离散分析 <sup>a,b</sup>			
	熵	集中度	df
模型	4.372	2.716	4
残差	773.727	437.635	1976
总计	778.098	440.351	1980
a. 模型：多项 Logit			
b. 设计：常量 + 相信死后有来世 + 相信死后有来世 * 种族 + 相信死后有来世 * 性别			

熵本是一个物理学概念，它是用来描述系统是否处于平衡状态的一种测度标准。它的一般计算公式为

$$H(\xi)=-\sum_{i=1}^r \hat{p}_i \ln \hat{p}_i$$

式中,  $i$  是类别指示符, 用正整数表示,  $i=1, \cdots, r$ ;  $r$  是类别数;  $\hat{p}_i$  是第  $i$  类出现的期望概率;  $\xi$  是随机变量。在给定  $\sum_{i=1}^k \hat{p}_i=1$  的条件约束下, 熵用来衡量给定分布与均匀分布的接近程度。

熵的值越高, 说明离散程度越大, 给定分布越接近于均匀分布, 越处于平衡状态。

在集中度统计量的计算中, 用到另一个描述分类变量离散程度的 Gini-Simpson (基尼-辛卜生) 指数, 简称 G-S 指数。其计算公式为

$$\text{G-S}(\xi)=1-\sum_{i=1}^r \hat{p}_i^2$$

G-S 值越小, 随机变量的分布越集中; 反之, 随机变量的分布越离散。

在表 11-103 中, 总熵=模型熵+残差熵, 与方差分析中总偏差平方和的分解公式很相似。下面用  $S(B)$  表示总熵, 用  $S(A)$  表示模型熵, 用  $S(B|A)$  表示残差熵, 来说明这些值是如何计算得到的。

总的平衡式为

$$S(B)=S(A)+S(B|A)$$

总熵为

$$S(B)=-N \sum_{i=1}^r S_i(B)$$

式中,  $S_i(B)=\hat{p}_i \ln \hat{p}_i$ , 显然, 它等于总观测频数乘以  $H$ 。

残差熵为

$$S(B|A)=-\sum_{j=1}^c N_j \sum_{i=1}^r S_{ij}(B|A)$$

式中,  $S_{ij}(B|A)=\hat{p}_{ij} \ln \hat{p}_{ij}$ 。

因此, 模型熵可用下式计算:

$$S(A)=S(B)-S(B|A)$$

在集中性表中, 同样用  $S(B)$  表示总集中度, 用  $S(A)$  表示模型集中度, 用  $S(B|A)$  表示残差集中, 也存在总的平衡式

$$S(B)=S(A)+S(B|A)$$

总集中度的计算公式为

$$S(B)=N \left( 1-\sum_{i=1}^r \hat{p}_i^2 \right)$$

也就是, 它等于总观测频数乘以 G-S。

$$S(B|A)=\sum_{j=1}^c N_j \left( 1-\sum_{i=1}^r \hat{p}_i^2 \right)$$

式中,  $A$  为通用的分类自变量(解释变量), 其类别用一些整数表示;  $B$  为通用的分类因变量(响应变量), 其类别用一些整数表示;  $r$  为  $B$  的分类数,  $r \geq 1$ ;  $c$  为  $A$  的分类数,  $c \geq 1$ ;  $i$  为  $B$  类别的通用指示符,  $i=1, \dots, r$ ;  $j$  为  $A$  类别的通用指示符,  $j=1, \dots, c$ ;  $N_j$  为  $A$  的第  $j$  个组的边际

总数,  $N_j = \sum_{i=1}^r n_{ij}$ ;  $n_{ij}$  为  $B$  的第  $i$  个响应与  $A$  的第  $j$  组对应单元格中的观测频数;  $N$  为总观测频数,  $N = \sum_{j=1}^c \sum_{i=1}^r n_{ij}$ 。

表 11-103 中, 模型熵值与模型集中度值都较小, 说明在因变量的总变异中, 由模型所能解释的部分很少, 主要是由模型以外的其他因素引起了这种变异。

对 Logit 对数线性模型而言, 没有一个精确的公式可以计算线性回归中的  $R^2$  统计量, 只能近似用计算的关联度  $R$  来替代。关联度  $R$  的计算公式为

$$R = S(A) / S(B)$$

显然,  $R$  在  $0 \sim 1$  之间取值。熵、集中度的关联度统计量值越大, 越接近于 1, 表明由模型解释的离散性越大; 反之, 由模型解释的离散性越小。

根据上面的公式, 可以计算熵的关联度为  $4.372/778.098=0.005619$ , 集中度的关联度为  $2.716/440.351=0.006168$ , 这就是表 11-104 中熵和集中度的关联度的结果。由于模型的关联度很小, 因此, 有必要加入其他一些自变量来改善模型的关联度。

表 11-105 列出了观测频数、期望频数、残差等统计量值。期望频数的“%”值是各组合条件下的期望概率值, 利用本表中值, 根据模型总熵、残差熵等计算公式, 就可以得到表 11-103 中的各个值。感兴趣的读者, 不妨自行验证一下。

表 11-104 关联度<sup>a, b</sup>

熵	.006
集中度	.006

模型: 多项 Logit 设计: 常量 + 相信死后有来世 + 相信死后有来世 \* 种族 + 相信死后有来世 \* 性别

表 11-105 观测频数、期望频数、残差表

单元计数和残差<sup>a, b</sup>

种族	性别	相信死后有来世	观测		期望的		残差	标准化残差	调整残差	偏差
			计数	%	计数	%				
黑人	女性	是	64	72.7%	62.247	70.7%	1.753	.411	.739	1.885
		不确定	9	10.2%	8.816	10.0%	.184	.065	.116	.610
		否	15	17.0%	16.937	19.2%	-1.937	-.524	-.926	-1.909
	男性	是	25	58.1%	26.753	62.2%	-1.753	-.551	-.739	-1.841
		不确定	5	11.6%	5.184	12.1%	-.184	-.086	-.116	-.601
		否	13	30.2%	11.063	25.7%	1.937	.676	.926	2.048
白人	女性	是	371	75.1%	372.753	75.5%	-1.753	-.183	-.739	-1.870
		不确定	49	9.9%	49.184	10.0%	-.184	-.028	-.116	-.606
		否	74	15.0%	72.063	14.6%	1.937	.247	.925	1.981
	男性	是	250	68.3%	248.247	67.8%	1.753	.196	.739	1.876
		不确定	45	12.3%	44.816	12.2%	.184	.029	.116	.607
		否	71	19.4%	72.937	19.9%	-1.937	-.253	-.925	-1.955

a. 模型: 多项 Logit

b. 设计: 常量 + 相信死后有来世 + 相信死后有来世 \* 种族 + 相信死后有来世 \* 性别

表 11-106 所示为模型中所有可能参数的估计值, 及其  $Z$  检验。 $Z$  = 估计/标准误。它有渐近的正态分布。因此, 在  $Z$  值的基础上, 可得到其  $p$  值。这里所作的假设为各参数的估计 = 0。

表 11-106 参数估计表

参数估计 <sup>a,d</sup>							
参数		估计	标准误	Z	Sig.	95% 置信区间	
						下限	上限
常量	[种族 = 0] * [性别 = 0]	2.830 <sup>a</sup>					
	[种族 = 0] * [性别 = 1]	2.404 <sup>a</sup>					
	[种族 = 1] * [性别 = 0]	4.278 <sup>a</sup>					
	[种族 = 1] * [性别 = 1]	4.290 <sup>a</sup>					
[相信死后有来世 = 1]		1.225	.128	9.561	.000	.974	1.476
[相信死后有来世 = 2]		-.487	.183	-2.659	.008	-.846	-.128
[相信死后有来世 = 3]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 1] * [种族 = 0]		-.342	.237	-1.442	.149	-.806	.123
[相信死后有来世 = 1] * [种族 = 1]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 2] * [种族 = 0]		-.271	.354	-.765	.444	-.965	.423
[相信死后有来世 = 2] * [种族 = 1]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 3] * [种族 = 0]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 3] * [种族 = 1]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 1] * [性别 = 0]		.419	.171	2.444	.015	.083	.754
[相信死后有来世 = 1] * [性别 = 1]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 2] * [性别 = 0]		.105	.247	.426	.670	-.378	.588
[相信死后有来世 = 2] * [性别 = 1]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 3] * [性别 = 0]		0 <sup>b</sup>	.	.	.	.	.
[相信死后有来世 = 3] * [性别 = 1]		0 <sup>b</sup>	.	.	.	.	.

- a. 在多项式假设中常量不作为参数使用。因此不计算它们的标准误差。
- b. 此参数为冗余参数，因此将被设为零。
- c. 模型：多项 Logit
- d. 设计：常量 + 相信死后有来世 + 相信死后有来世 \* 种族 + 相信死后有来世 \* 性别

在本题中，重点要关注“相信死后有来世”与“种族”，及“相信死后有来世”与“性别”之间的交互项的参数估计的检验结果。当 $p$ 值小于 0.05 时，则拒绝交互作用项的参数估计为 0 的假设，说明这两个变量之间存在关联关系，否则这两个变量之间是独立的。从表中可见，只有“相信死后有来世”与“性别”之间的 $p$ 值小于 0.05，表明“相信死后有来世”与“性别”之间存在关联关系，而“相信死后有来世”与“种族”之间是独立的。

在交互作用项中，有显著性意义的参数估计的值为 0.419( $p = 0.015$ )>0，它对应于“性别”=0，即女性，软件默认的基准类别为 1(1>0)，即男性，而“相信死后有来世”=1 的类别为“是”。因此，女性“相信死后有来世”的估计概率要比男性“相信死后有来世”的概率要高。

事实上，表中的效应参数描述的是与基准类别的优势比的对数。根据 Logit 对数线性模型建模的规则可知，在本例中因变量  $y$ (相信死后有来世)的基准类为 3，即“否”。因此，0.419 也是给定“种族”、“性别”和响应类别是与否之间的优势比的条件对数。

对于女性来说，在确定“种族”后，“相信死后有来世”与“不相信死后有来世”的概率，是男性的  $e^{0.419}$  倍，也就是 1.52 倍。

同样，对于黑人而言，在确定“性别”后，黑人“相信死后有来世”与“不相信死后有来世”的概率是白人的  $e^{-0.342}$  倍，也就是 0.71 倍，但这没有统计学上的显著性意义。

利用表 11-106 中提供的参数估计的数据，令  $s = 1$  为女性， $s = 0$  为男性， $r = 1$  为黑人， $r = 0$  为白人，则可以得到 Logit 对数线性模型

$$\ln \frac{P(y=1)}{P(y=3)} = 1.225 + 0.419s - 0.342r$$
$$\ln \frac{P(y=2)}{P(y=3)} = -0.487 + 0.105s - 0.271r$$

由此，还可以得到

$$\begin{aligned} \ln \frac{P(y=1)}{P(y=2)} &= \ln \frac{\frac{P(y=1)}{P(y=3)}}{\frac{P(y=2)}{P(y=3)}} = \ln \frac{P(y=1)}{P(y=3)} - \ln \frac{P(y=2)}{P(y=3)} \\ &= 1.225 + 0.419s - 0.342r - (-0.487 + 0.105s - 0.271r) = 1.71 + 0.314s - 0.071r \end{aligned}$$

有了这些模型，加上在因变量各类上的累计概率和为 1 的条件约束，就可计算出表 11-105 中期望的“%”值。

如果在模型中不包含截距，与自变量最后一类有关的参数将不再是冗余的。

表 11-107 与表 11-108 分别给出了参数估计的相关矩阵与协方差矩阵。

表 11-107 参数估计的相关性表

参数估计的相关性 <sup>a,b,c</sup>						
	[相信死后有来世 = 1]	[相信死后有来世 = 2]	[相信死后有来世 = 1] * [种族 = 0]	[相信死后有来世 = 2] * [种族 = 0]	[相信死后有来世 = 1] * [性别 = 0]	[相信死后有来世 = 2] * [性别 = 0]
[相信死后有来世 = 1]	1	.539	-.226	-.124	-.692	-.367
[相信死后有来世 = 2]	.539	1	-.129	-.219	-.370	-.691
[相信死后有来世 = 1] * [种族 = 0]	-.226	-.129	1	.512	-.081	-.045
[相信死后有来世 = 2] * [种族 = 0]	-.124	-.219	.512	1	-.042	-.075
[相信死后有来世 = 1] * [性别 = 0]	-.692	-.370	-.081	-.042	1	.553
[相信死后有来世 = 2] * [性别 = 0]	-.367	-.691	-.045	-.075	.553	1

- a. 模型：多项 Logit  
b. 设计：常量 + 相信死后有来世 + 相信死后有来世 \* 种族 + 相信死后有来世 \* 性别  
c. 未显示常量和冗余的参数。

表 11-108 参数估计的协方差矩阵表

参数估计的协方差 <sup>a,b,c</sup>						
	[相信死后有来世 = 1]	[相信死后有来世 = 2]	[相信死后有来世 = 1] * [种族 = 0]	[相信死后有来世 = 2] * [种族 = 0]	[相信死后有来世 = 1] * [性别 = 0]	[相信死后有来世 = 2] * [性别 = 0]
[相信死后有来世 = 1]	.016	.013	-.007	-.006	-.015	-.012
[相信死后有来世 = 2]	.013	.034	-.006	-.014	-.012	-.031
[相信死后有来世 = 1] * [种族 = 0]	-.007	-.006	.056	.043	-.003	-.003
[相信死后有来世 = 2] * [种族 = 0]	-.006	-.014	.043	.125	-.003	-.007
[相信死后有来世 = 1] * [性别 = 0]	-.015	-.012	-.003	-.003	.029	.023
[相信死后有来世 = 2] * [性别 = 0]	-.012	-.031	-.003	-.007	.023	.061

- a. 模型：多项 Logit  
b. 设计：常量 + 相信死后有来世 + 相信死后有来世 \* 种族 + 相信死后有来世 \* 性别  
c. 未显示常量和冗余的参数。

图 11-98、图 11-99 及图 11-100 所示是 3 张诊断图。图 11-98 所示为 12 个单元格的观测频数、期望频数和校正残差两两对应的散点图，可见 12 个散点分布很随机。图 11-99、图 11-100 所示的 Q-Q 图中，散点基本分布在直线上，及在 0 线周围随机分布，说明残差近似服从正态分布。

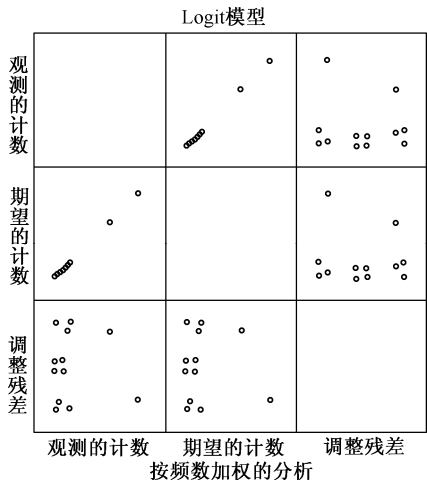


图 11-98 观测频数、期望频数和调整残差两两对应的散点图

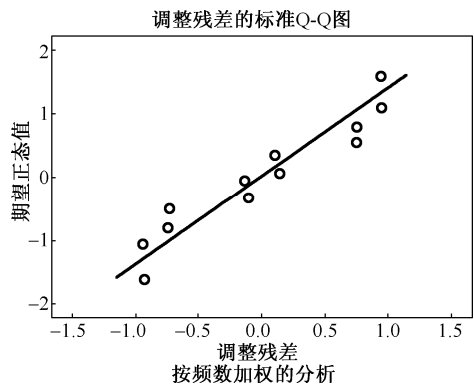


图 11-99 校正残差的正态 Q-Q 图

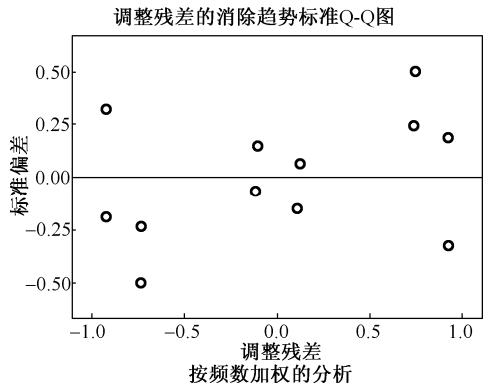


图 11-100 校正残差的去势正态 Q-Q 图

11.11.4 模型选择对数线性回归分析

模型选择过程使用迭代比例拟合算法对多维列联表拟合分层对数线性模型。在研究人员对分类变量之间的因果关系不甚了解，分不清哪个是因变量，哪个是自变量，或只是想了解分类变量之间是否可能存在关联关系时，可以首选分层对数线性模型，用它可以找出关联的分类变量。

在用分层对数线性模型建模时，可以使用强制输入法，也可以使用向后剔除法。向后剔除法是从饱和模型入手的，它从高阶交互项开始逐步剔除无意义的参数，直到得到最佳简约模型为止。

向后剔除法与多元线性回归中的逐步回归有很多相似之处，它能对进出模型的变量进行自动筛选，这在高维列联表上进行联合分析时，可以节省大量工作时间，因而很受广大研究人员的欢迎。但需要注意的是，在模型选择对数线性回归过程中的向后剔除法，是当所有  $K+1$  阶的交互作用项都无显著性意义，全部剔除出模型后，才去考虑  $K$  阶交互作用项是否被剔除的问题。

饱和模型会给所有单元格加上 0.5，目的是为了 避免抽样 0 的出现。另外，只有选用饱和模型，才可以要求输出参数估计值和关联性检验。

注意：利用模型选择过程来获取模型中需要哪些项的信息，再根据需要使用一般对数线性分析或 Logit 对数线性分析来继续评估模型。

1. 对数据的要求

- (1) 因子变量必须是分类变量。
- (2) 要分析的变量必须是数值型变量。对于字符型的分类变量，在分析之前需要采用前面介绍过的自动重新编码的方式将其重新编码为数值型变量。如果数值型变量中有空类别，则使用“重新编码”创建连续的整数值。
- (3) 要避免指定多个多水平的分类变量。因为这有可能导致多个单元格中只有少量的观察值出现，从而无法使用卡方检验。

2. 模型选择对数线性回归分析过程

- (1) 按【分析→对数线性模型→模型选择】顺序，打开【模型选择对数线性分析】对话框，见图 11-101。
- (2) 选定模型中需要的各种变量。从左侧的源变量框中选择多个分类变量作为因子变量进入【因子】框。注意，其中大多数变量应为二分变量。  
对每个选定的变量，需要进行类别取值范围的定义，方法为在【因子】框中选定变量，单击【定义范围】按钮，在弹出的【对数线性分析：定义范围】对话框中(见图 11-102)的【最小值】框中输入定义类别中的最小整数值，在【最大值】框中输入定义类别中的最大整数值，最小值必须小于最大值。单击【继续】按钮，返回【模型选择对数线性分析】对话框。重复这个定义过程，直到将所有因子变量都定义完毕。变量定义完成后，在其变量名后面的括号中会出现所定义的最小和最大值。
- 注意：在最小值和最大值定义的范围之外的个案，将不用于建模中。
- 从左侧的源变量框中选择一个权重变量，将其移入到【单元格权重】框中。
- (3) 选择建模方法。在【建立模型】栏中，提供了两种可选的变量进出模型的方法，剔除变量的显著性水平，及迭代比例拟合算法的收敛标准。



图 11-101 【模型选择对数线性分析】对话框

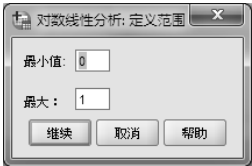


图 11-102 【对数线性分析：定义范围】对话框

- ① 【使用向后排除法】(向后剔除法)。从饱和模型中逐渐从最高阶交互项中剔除无意义的参数，直到得到最佳简约模型为止。
  - ② 【一步进入法】。如果建立的是不饱和模型，则选择此项，它将所有变量一次性强制选入模型。
- 另外，在【最多步骤数】框中，需输入最多迭代的次数，系统默认值为 10。在【删除概率】框中，输入从模型中剔除变量的显著性水平的标准，系统默认值为 0.05，大于此概率值的变量将被从模型中剔除。

(4) 定义模型。单击【模型】按钮，打开如图 11-103 所示的【对数线性分析：模型】对话框。在【指定模型】栏中，有两个选项：

- ①【饱和】。在模型中包含所有因子的主效应以及所有因子间的交互作用。
- ②【设定】。用户可为不饱和模型自定义生成类。

在左侧【因子】列表框中，选定要进入模型的变量组合，在【构建项】下拉列表的 6 个选项中选择 1 个变量间交互作用的方式，单击右移箭头，则在【生成类】框中列出所需要的生成类。

在【生成类】框中列出的是因子出现的最高阶项的列表。在分层模型中将包括定义的生成类及其比他低阶的所有相关的项。例如，在左侧的【因子】列表中，选择变量  $A$ 、 $B$  和  $C$ ，然后在【构建项】下拉列表中选择【交互】。在【生成类】框中，出现  $A*B*C$  项，则在生成的模型中将包含指定的三阶交互  $A*B*C$ ，二阶交互  $A*B$ 、 $A*C$  和  $B*C$ ，以及  $A$ 、 $B$  和  $C$  的主效应。因此，不要在生成类中指定低阶的相关性。

按【继续】按钮，返回【模型选择对数线性分析】对话框。

(5) 定义输出结果中的显示项。单击【选项】按钮，打开如图 11-104 所示的【对数线性分析：选项】对话框，可以选择输出频数、残差等统计量、统计图，饱和模型状态下的参数估计、关联表等，还可以选择模型拟合过程中的迭代收敛的标准。

- ①【输出】栏。
  - 【频率】(应为频数)。输出频数表。此项为系统默认选项。
  - 【残差】。在输出表中包含残差项信息。此项为系统默认选项。
- 在饱和模型中，观测频数和期望频数相同，残差为 0。
- ②【图】栏。在饱和模型下，该栏无效。
  - 【残差图】。输出残差图。
  - 【正态概率】。输出正态概率图。



图 11-103 【对数线性分析：模型】对话框



图 11-104 【对数线性分析：选项】对话框

这些图形可帮助确定模型与数据的拟合度。

- ③【显示饱和模型】栏。在不饱和模型下，本栏无效。
  - 【参数估计】。显示饱和模型的参数估计表，这些值有助于确定可从模型中删除哪一项。
  - 【相关表】(关联表)。显示内含偏关联检验的关联表。需要注意的是，如果模型中选择了多个因子，则选择本选项需要进行大量的计算。
- ④【模型标准】栏。需指定使用迭代比例拟合算法来获取参数估计值中的最大迭代次数、收敛性或 Delta。



- 【最大迭代】次数。系统默认值为 20。
- 【收敛性】标准。可在其下拉菜单中加以选择。系统默认值为【默认】。
- 【Delta】值。用来设置饱和模型的校正系数。系统默认值为 0.5。

单击【继续】按钮，返回【模型选择对数线性分析】对话框。

单击【确定】按钮，则在输出窗中得到运行结果。

3. 模型选择对数线性回归实例分析

【例 16】仍以【例 15】中的数据为例，使用模型选择过程来分析“性别”、“种族”与“相信死后有来世”之间的关联。

操作步骤如下：

(1) 打开数据文件 data11-13。按【分析→对数线性模型→模型选择】顺序，打开【模型选择对数线性分析】对话框。

(2) 从左侧的源变量框中选择“性别”、“种族”与“相信死后有来世”分类变量作为因子变量进入【因子】框。

在【因子】框中，选定“性别”变量，单击【定义范围】按钮，在弹出的【对数线性分析：定义范围】对话框中的【最小值】框中输入 0，在【最大值】框中输入 1。单击【继续】按钮，返回【模型选择对数线性分析】对话框。

选定“种族”变量，单击【定义范围】按钮，在弹出的【对数线性分析：定义范围】对话框中的【最小值】框中输入“0”，在最大值框中输入“1”。单击【继续】按钮，返回【模型选择对数线性分析】对话框。

选定“相信死后有来世”变量，单击【对数线性分析：定义范围】按钮，在弹出的【定义范围】对话框中的【最小值】框中输入“1”，在【最大值】框中输入“3”。单击【继续】按钮，返回【模型选择对数线性分析】对话框。

对话框中的其他选项均选用系统默认选项。这意味着，将从饱和模型出发，使用向后排除法(向后剔除法)来建立模型。这可以在输出中要求显示参数估计表。

由于本例已选择饱和模型，这是【指定模型】框中的默认选项，可不作任何选择。

(3) 单击【选项】按钮，打开【对数线性分析：选项】对话框。

在【显示饱和模型】栏中，选择【参数估计】和【相关表】(关联表)选项，其他保持系统默认选项。

单击【继续】按钮，返回【模型选择对数线性分析】对话框。

(4) 单击【确定】按钮，则在输出窗中得到表 11-109～表 11-119 的结果。

(5) 结果分析。

表 11-109 数据信息

数据信息		N
个案	有效	12
	超出范围 <sup>a</sup>	0
	缺失	0
	加权有效	991
类别	种族	3

a. 由于超过因子值范围，  
个案被拒绝。

表 11-110 收敛信息

收敛信息	
生成类	种族*性别*相信死后有来世
迭代数	1
“观测边际”与“拟合边际”之间的最大差异	.000
收敛性准则	.371

表 11-109 所示为参与拟合模型数据的基本信息，共有 12 个单元格频数，没有超出定义范围之外的数据，也没有缺失值，样本量为 991，有 3 个变量参与建模。

模型中设定的生成类为“种族\*性别\*相信死后有来世”，由于模型中只有 3 个变量，因此，设定的是饱和模型。迭代次数为 1，收敛性准则为 0.371，表明只用 1 次迭代运算就达到 0.5 的收敛标准，完成饱和模型的拟合。观测频数边际和与拟合频数的边际和之间的最大差异为 0，表示两个边际和是相同的。

表 11-111 所示为原始的观测频数与用饱和模型计算得到的期望频数，以及它们之间的残差。在饱和模型状态下，期望频数总是与观测频数相同，因此残差必为 0。换言之，该表在饱和模型状态下没有必要看。

表 11-111 观测频数、期望频数、残差

单元计数和残差								
种族	性别	相信死后有来世	观测		期望		残差	标准残差
			计数 <sup>a</sup>	%	计数	%		
黑人	女性	是	64.500	6.5%	64.500	6.5%	.000	.000
		不确定	9.500	1.0%	9.500	1.0%	.000	.000
		否	15.500	1.6%	15.500	1.6%	.000	.000
	男性	是	25.500	2.6%	25.500	2.6%	.000	.000
		不确定	5.500	0.6%	5.500	0.6%	.000	.000
		否	13.500	1.4%	13.500	1.4%	.000	.000
白人	女性	是	371.500	37.5%	371.500	37.5%	.000	.000
		不确定	49.500	5.0%	49.500	5.0%	.000	.000
		否	74.500	7.5%	74.500	7.5%	.000	.000
	男性	是	250.500	25.3%	250.500	25.3%	.000	.000
		不确定	45.500	4.6%	45.500	4.6%	.000	.000
		否	71.500	7.2%	71.500	7.2%	.000	.000

a. 对于饱和模型，.500 已添加至所有观测单元格中。

同样，在饱和模型状态下，其拟合优度检验表(见表 11-112)也没有任何意义。两个检验统计量的计算结果总为 0，自由度也总为 0，无法给出检验的概率值，因此，系统将其设为缺失值。

表 11-112 拟合优度检验

拟合优度检验			
	卡方	df	Sig.
似然比	.000	0	.
Pearson	.000	0	.

表 11-113 所示为 K 阶及更高阶效应检验结果。其第二栏，即“K-Way 和高阶效果”栏，不是很好理解，主要是由于汉化时未能将该句子的意思完整正确地转译过来。原文为：Tests that K-way and higher order effects are zero，其确切含义应为：K 阶及更高阶效应等于 0 的检验。同样，第三栏应为：K 阶效应等于 0 的检验。

表 11-113 K 阶及更高阶效应检验

K-Way 和高阶效果							
K	df	似然比		Pearson		迭代数	
		卡方	Sig.	卡方	Sig.		
K-Way 和高阶效果 <sup>a</sup>	1	11	1265.541	.000	1676.669	.000	0
	2	7	14.141	.049	13.604	.059	2
	3	2	.854	.653	.861	.650	3
K-way 效果 <sup>b</sup>	1	4	1251.400	.000	1663.065	.000	0
	2	5	13.287	.021	12.743	.026	0
	3	2	.854	.653	.861	.650	0

a. 检验 k-way 和高阶效果是否为零。

b. 检验 k-way 效果是否为零。

在指定饱和模型条件下,可以要求作效应的偏关联分析。设  $\chi^2(k)$  表示包含主效应及  $K$  阶交互项模型的卡方值。第  $K$  阶交互作用的显著性检验可基于  $\chi^2(k-1) - \chi^2(k)$  来进行。而自由度则可通过减去相应模型的自由度来获取。

由此可知:

第二栏中第一个检验是检验模型中一阶(主效应)及以上阶的交互效应为 0 的假设,  $p = 0.000 < 0.05$ , 有充分的证据可以拒绝原假设, 因此, 主效应及其交互效应有统计学意义。也就是“性别”、“种族”、“相信死后有来世”、“性别\*种族”、“种族\*相信死后有来世”、“性别\*相信死后有来世”、“性别\*种族\*相信死后有来世”中, 至少有一项交互效应有统计学意义。

第二栏中第二个检验是检验模型中二阶及以上阶的交互效应为 0 的假设,  $p = 0.049 < 0.05$ , 有充分的证据可以拒绝原假设, 因此, 二阶及以上阶的交互效应有统计学意义。也就是“性别\*种族、种族\*相信死后有来世”、“性别\*相信死后有来世、性别\*种族\*相信死后有来世”中, 至少有一项交互效应有统计学意义。

第二栏中第三个检验是检验模型中三阶及以上阶的交互效应为 0 的假设,  $p = 0.653 > 0.05$ , 没有充分的证据可以拒绝原假设, 因此, 三阶及以上阶的交互效应没有统计学意义。也就是三阶“性别\*种族\*相信死后有来世”的交互作用没有统计学意义。

用第二栏中第一个似然比卡方值检验的结果减去第二栏中第二个似然比卡方值检验的结果, 可以得到熵之差  $\Delta_1 G = 1265.541 - 14.141 = 1251.400$ ,  $df_1 = 11 - 7 = 4$ , 这也就是表中第三栏第一个似然比卡方值检验的结果。它是检验模型中一阶交互效应(即主效应)为 0 的假设,  $p = 0.000 < 0.05$ , 有充分的证据可以拒绝原假设, 因此, 主效应有统计学意义, 即“性别”、“种族”、“相信死后有来世”中至少有一个的效应不为 0。

注意: 如果用 Pearson 卡方值来分析也同样有上面的结论, 下同。

同理, 用第二栏中第二个似然比卡方值检验的结果减去第二栏中第三个似然比卡方值检验的结果, 可以得到熵之差  $\Delta_1 G = 14.141 - 0.854 = 13.287$ ,  $df_1 = 7 - 2 = 5$ , 这也就是表中第三栏第二个似然比卡方值检验的结果。它是检验模型中二阶交互效应(即主效应)为 0 的假设,  $p = 0.021 < 0.05$ , 有充分的证据可以拒绝原假设, 因此, 二阶交互效应有统计学意义, 即“性别\*种族”、“种族\*相信死后有来世”、“性别\*相信死后有来世”中至少有一个的效应不为 0。

由于本例中参与模型拟合的变量只有“性别”、“种族”、“相信死后有来世”3 个, 最高的交互作用项只有一个, 即“性别\*种族\*相信死后有来世”。因此, 第二栏中第三个检验就是在检验“性别\*种族\*相信死后有来世”的交互效应=0, 它同第三栏中第三个检验是等价的, 所以, 它们有相同的检验结果。

由上面的分析可以得到: “性别”、“种族”、“相信死后有来世”3 个变量之间有主效应和二阶交互效应存在。

表 11-114 所示为有显著性意义的主效应和二阶效应的各项名称。由该表可知, 在二阶交互项中, 有显著性意义的有两项, 分别为“性别\*种族”( $p = 0.026 < 0.05$ )和“种族\*相信死后有来世”( $p = 0.027 < 0.05$ ), 主效应中“种族”( $p = 0.00 < 0.05$ )有显著性意义。

表 11-115 所示为在饱和模型条件下, 部分主效应、交互效应的参数估计及显著性检验结果。“种族”的两个参数估计均有显著性意义( $p = 0.000 < 0.05$ ), “种族\*相信死后有来世”的第一个参数估计有显著性意义( $p = 0.033 < 0.05$ )。

表 11-114 偏关联检验

偏关联				
效果	df	偏卡方	Sig.	迭代数
种族*性别	1	4.987	.026	2
种族*相信死后有来世	2	7.193	.027	2
种族	2	621.253	.000	2

表 11-115 参数估计值

参数估计值							
效果	参数	估计	标准误	Z	Sig.	95% 置信区间	
						下限	上限
种族*性别*相信死后有来世	1	.042	.072	.591	.555	-.098	.183
	2	.024	.104	.236	.813	-.179	.227
种族*性别	1	.091	.062	1.465	.143	-.031	.213
种族*相信死后有来世	1	.153	.072	2.131	.033	.012	.293
	2	-.020	.104	-.193	.847	-.223	.183
种族	1	1.008	.072	14.049	.000	.867	1.148
	2	-.785	.104	-7.579	.000	-.988	-.582

表 11-116 中的步骤 0，表示从这里正式开始分析，由生成类可知，分析是从饱和模型开始的，此时卡方值为 0。删除最高阶交互项“性别\*种族\*相信死后有来世”后， $p = 0.653 > 0.05$ ，模型拟合效果无显著性变化，说明三阶交互效应不存在。

表 11-116 逐步筛选过程摘要

步骤摘要						
步骤 <sup>a</sup>		效果	卡方 <sup>c</sup>	df	Sig.	迭代数
0	生成类 <sup>b</sup>	种族*性别*相信死后有来世	.000	0	.	
	已删除的效果	1 种族*性别*相信死后有来世	.854	2	.653	3
1	生成类 <sup>b</sup>	种族*性别, 种族*相信死后有来世, 性别*相信死后有来世	.854	2	.653	
	已删除的效果	1 种族*性别	4.987	1	.026	2
		2 种族*相信死后有来世	1.994	2	.369	2
		3 性别*相信死后有来世	7.193	2	.027	2
2	生成类 <sup>b</sup>	种族*性别, 性别*相信死后有来世	2.848	4	.584	
	已删除的效果	1 种族*性别	4.543	1	.033	2
		2 性别*相信死后有来世	6.749	2	.034	2
3	生成类 <sup>b</sup>	种族*性别, 性别*相信死后有来世	2.848	4	.584	

- a. 在每一步骤中，如果最大显著性水平大于 .050，则删除含有“似然比更改”的最大显著性水平的效果。
- b. 在步骤 0 之后，将在每一步骤显示最佳模型的统计量。
- c. 对于“已删除的效果”，从模型中删除该效果之后，这是卡方中的更改。

拟合的步骤如下。

步骤 1：模型中当前有 3 个二阶交互项时，当从模型中逐一删去一个交互项时，发现删去“性别\*种族” ( $p = 0.026 < 0.05$ ) 和“种族\*相信死后有来世” ( $p = 0.027 < 0.05$ ) 时，模型拟合效

果有显著性变化,说明“性别\*种族”、“性别\*相信死后有来世”之间有交互作用,而“种族\*相信死后有来世”之间没有交互作用( $p = 0.369 < 0.05$ )。

步骤 2: 在步骤 1 的基础上,继续逐一剔除不显著的交互项,由于剩下的两个交互项检验的  $p$  值均小于 0.05,剔除后模型的拟合效果将有显著变化,故不能再剔除。

步骤 3: 得到最终的拟合模型,其中包括“性别”、“种族”、“相信死后有来世”、“性别\*种族”、“性别\*相信死后有来世”5 项。

表 11-117 给出了模型拟合过程步骤 0 中的收敛信息。

表 11-117 收敛信息<sup>a</sup>

生成类	种族*性别, 性别*相信死后有来世
迭代数	0
“观测边际”与“拟合边际”之间的最大差异	.000
收敛性准则	.371

最终模型的统计量在“反向消除”之后。

表 11-118 给出了用“性别”、“种族”、“相信死后有来世”、“性别\*种族”、“性别\*相信死后有来世”5 项建模后,得到的期望频数与观测频数的对照结果。

表 11-119 所示为对用“性别”、“种族”、“相信死后有来世”、“性别\*种族”、“性别\*相信死后有来世”5 项所建模型进行的拟合优度检验结果。如果显著性的值很小(小于 0.05),则模型不能充分拟合数据。在本例中,因为两个检验方法的  $p$  值均大于 0.05,没有理由去拒绝原假设,所以模型拟合效果不错。

表 11-118 观测频数、期望频数及残差

			单元计数和残差							
种族	性别	相信死后有来世	观测		期望		残差	标准残差		
			计数	%	计数	%				
黑人	女性	是	64.000	6.5%	65.773	6.6%	-1.773	-.219		
		不确定	9.000	0.9%	8.770	0.9%	.230	.078		
		否	15.000	1.5%	13.457	1.4%	1.543	.421		
	男性	是	25.000	2.5%	28.912	2.9%	-3.912	-.728		
		不确定	5.000	0.5%	5.257	0.5%	-.257	-.112		
		否	13.000	1.3%	8.831	0.9%	4.169	1.403		
白人	女性	是	371.000	37.4%	369.227	37.3%	1.773	.092		
		不确定	49.000	4.9%	49.230	5.0%	-.230	-.033		
		否	74.000	7.5%	75.543	7.6%	-1.543	-.178		
	男性	是	250.000	25.2%	246.088	24.8%	3.912	.249		
		不确定	45.000	4.5%	44.743	4.5%	.257	.038		
		否	71.000	7.2%	75.169	7.6%	-4.169	-.481		

表 11-119 拟合优度检验

拟合优度检验			
	卡方	df	Sig.
似然比	2.848	4	.584
Pearson	3.076	4	.545

## 习 题 11

1. 数据文件 data11-14 是某企业 1987—1998 年的经济效益、科研人员、科研经费的统计数据。假定 1999 年该企业科研人员 61 名、科研经费 40 万元，试预测 1999 年该企业的经济效益。

2. 某商场 1989—1998 年的商品流通费用与商品零售额资料如数据文件 data11-15 所示。若 1999 年该商场商品零售额为 36.33 亿元，试预测 1999 年该商场商品流通费用。

3. 数据文件 data11-16 是 R.Norell 进行的一项用电流刺激农场动物的试验数据，其目的是为了求得一成的牲畜对高压电流有反应的临界值。在新农场选址时，要求高压线的辐射电流低于临界值，如果超过，则需要重新选址。试求出临界电流值。

# 第 12 章 非参数检验

检验问题可划分为两大类：在已知总体分布的具体函数形式的前提下，只是其中若干个参数未知，则称这种检验问题为参数检验问题，否则称为非参数检验问题。在前面第 8 章提到的假设检验中，大都要求知道总体的分布，且检验与总体参数有关，称这类检验为参数假设检验。在实际问题中，人们往往不知道总体分布的类型，或者知之甚少，而又需要根据样本提供的信息对假设的总体分布进行检验。另外，在解决实际问题中遇到多维随机变量时，也需要对随机变量间是否具有独立性进行检验。这种和数据本身的总体分布无关的假设检验称为非参数假设检验。本章将介绍 SPSS 中用到的一些非参数假设检验方法。

SPSS 20 中进行非参数检验由主菜单的【分析】下拉菜单中的【非参数检验】菜单项导出。单击【分析】，用鼠标箭头指向【非参数检验】，显示子菜单，见图 12-1，其中包括【单样本】、【独立样本】、【相关样本】和【旧对话框】4 个子菜单。单击【旧对话框】，则弹出在先前 SPSS 版本中的非参数检验中用到的各种检验过程，见图 12-2，它们分别是【卡方】检验、【二项式】检验、【游程】检验、【1-样本 K-S】检验、【2 个独立样本】检验、【K 个独立样本】检验、【2 个相关样本】检验、【K 个相关样本】检验。

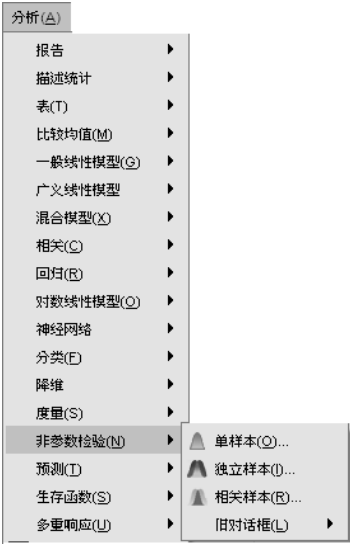


图 12-1 各种非参数检验



图 12-2 【旧对话框】中的各种非参数检验

在上述 8 种非参数检验方法中，前 4 种方法通常用来作分布的拟合优度检验，即检验样本所在的总体是否服从某个已知的理论分布；后 4 种方法通常用于分布位置检验，即检验样本所在的总体的分布位置或形状是否相同。

新版对旧版的检验过程进行了进一步的分类，将旧版中的前 4 项，即【卡方】检验、【二项式】检验、【游程】检验、【1-样本 K-S】检验，另加 Wilcoxon 符号秩检验合并成【单样本】检验，将【2 个独立样本】检验和【K 个独立样本】检验合并为独立样本检验，同样将【2 个相关样本】

检验和【K 个相关样本】检验合并为相关样本检验。这使得分类更加明确，但考虑到许多读者已习惯使用旧版来进行统计分析，接触新版界面还需要有一个过程，因此，本章仍以介绍旧版为主，而对新版界面的使用方法，将在第 12.9 节中作初步介绍。

## 12.1 卡 方 检 验

### 12.1.1 卡方检验的基本概念

在前面介绍的方法中，往往都事先假定总体服从正态分布，然后对其均值或方差作差异的显著性检验。但某个随机变量是否服从某种特定的分布是需要进行检验的。可以根据以往的经验或实际的观测数据的分布情况，推测总体可能服从某种分布函数  $F(x)$ ，利用这些样本数据来具体检验该总体分布函数是否真的就是  $F(x)$ 。卡方检验(Chi-Square Test)就是这样一种用来检验给定的概率值下数据来自同一总体的无效假设的方法。通常地，卡方检验可以用来对分类变量是二项或多项分布的总体作分布的一致性检验。

在卡方分布的一致性检验中，所作的原假设为

$$H_0: \text{类 } A_i \text{ 所占的比例为 } p_i = p_{i0} \quad (i = 1, 2, \cdots, r)$$

检验统计量为

$$\chi^2 = \sum_{i=1}^r \frac{(O_i - E_i)^2}{E_i}$$

式中， $A_i$  表示第  $i$  类； $O_i$  表示第  $i$  类中观察到的样本中实际出现的频数； $E_i$  表示第  $i$  类中理论上的期望频数， $E_i = np_{i0}$ ， $n$  为总的观察次数， $p_{i0}$  为第  $i$  类理论上出现的概率。

在原假设  $H_0$  成立时，每一类  $i$  中观察到的样本中实际出现的频数  $O_i$  ( $O_i = n_i$ ) 与理论上的期望频数  $E_i$  ( $E_i = np_{i0}$ ) 之间应相当接近，在原假设  $H_0$  成立时，上述统计量的渐近分布为  $\chi^2(r-1)$ 。在给定显著性水平  $\alpha$  下，当  $\chi^2 \geq \chi^2_{1-\alpha}(r-1)$ ，或  $P < 0.05$  时，拒绝原假设。

卡方检验适用于数值型有序或名义测度的分类变量。如果是字符型变量，则先要使用【转换】菜单中的【自动重新编码】过程将其转换成数值型变量。类别一般用整数表示。

在卡方检验中，假定数据来自于随机样本。每个类别的期望频数不小于 1，且期望频数小于 5 的类别数，不能超过总类别数的 20%。

### 12.1.2 卡方检验过程

- (1) 按【分析→非参数检验→旧对话框→卡方】顺序打开【卡方检验】对话框，见图 12-3。
- (2) 从左侧变量列表选择一个或多个需要进行检验的变量，单击向右箭头按钮，使变量移到【检验变量列表】框中。
- (3) 在【期望全距】(应为【期望范围】)栏内确定检验值的范围。在默认情况下，变量的各个截然不同的值被当作分类值。
  - ①【从数据中获取】。采用数据中的最小值和最大值所确定的范围。系统默认选项。
  - ②【使用指定的范围】。建立指定范围内的分类，只检验数据中一个子集的值，在【下限】和【上限】参数框中分别输入检验范围的下限和上限。输入的数值须为整数。数值超过这个指定范围的样品，不参与分析。



(4) 在【期望值】栏中指定期望值。

①【所有类别相等】。系统默认的检验是所有组对应的期望值都相同，这意味着要检验的是总体是否服从均匀分布。

②【值】。选定所要检验的是总体是否服从某个给定的分布，并在其右边的框中输入相应各组所对应的由给定分布所计算而得的期望值的百分比。该数值必须大于 0，并应同原分类次序相同的升序顺序保持一致。这一点非常重要。

每输入一个值后单击【添加】按钮，于是在它右边的框的底部便增加了刚输入的期望值百分比，一直到输完所有的期望值为止。如果在输入了错误数据后，选中错误数据，单击【删除】按钮即可删除；或者输入正确数据，单击【更改】按钮，则错误值被替换。

(5) 单击【选项】按钮，打开【卡方检验：选项】对话框，见图 12-4。

① 在【统计量】栏中选择输出统计量。

- 【描述性】。指定输出变量的均值、标准差、最大值、最小值、非缺失个体的数量。
- 【四分位数】。输出四分位数。

②【缺失值】栏中选择对缺失值的处理方式。

- 【按检验排除个案】。将参与对比中的缺失值排除。
  - 【按列表排除个案】(按记录排除个案)。剔除任何变量中所有含有缺失值的样品。
- 单击【继续】按钮，返回【卡方检验】主对话框。

(6) 单击【精确】按钮，打开【精确检验】对话框，见图 12-5。仅当购买了“精确检验选项”时，此功能才可使用。



图 12-3 【卡方检验】主对话框



图 12-4 【卡方检验：选项】对话框

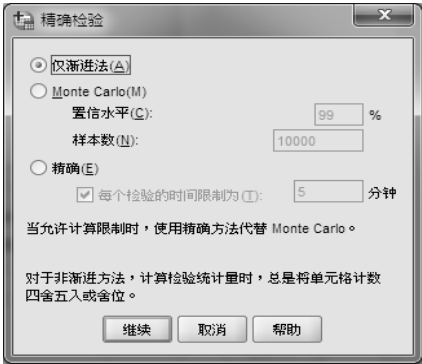


图 12-5 【精确检验】对话框

它提供了另外两种计算显著性水平的方法：精确法和蒙特卡洛(Monte Carlo)法。当数据不满足标准渐近法所必需的基本假定条件时，它们提供了一个获得精确结果的方法。在图 12-5 所示的对话框中，共有 3 个选项可选择：

①【仅(适用于)渐近法】。系统默认选项，就是上面提到的方法。它的使用条件为大样本

且各类的期望频数大于 5，满足这两个条件时可选择本项。

②【Monte Carlo】(蒙特卡洛)。本方法与【精确】检验都适用于数据集很小(如样本含量小于 30)，表格稀疏或不平衡，或样本含量小于 50 且出现小于 5 的期望频数时。选择此项，在【置信水平】框中输入置信水平，并在【样本数】框指定用于此近似法中的样品数量。要想复制结果，每次使用此法时要设置随机数种子。此方法比【精确】检验能更快得到结果。

③【精确】检验。适用条件同【Monte Carlo】法。选择本项，需要在【每个检验的时间限制为】后的框中输入最大限制时间。如果检验计算超过设置时限 30 min，建议使用 Monte Carlo 法。如果发现没有足够的内存空间，应首先关闭正在运行的其他应用软件，以节省内存供计算使用。如果还不能获得精确结果，应改用【Monte Carlo】法。

(7) 单击【确定】按钮，系统立即执行命令。或者单击【粘贴】按钮，在【语法】窗口中生成【卡方检验】命令程序。单击【运行】按钮，执行命令。

12.1.3 卡方检验分析实例

【例 1】 掷一颗六面体 300 次，结果见表 12-1，取变量名为“lmt”，用数值型数据 1、2、3、4、5、6 分别代表六面体各面的 6 个点，试问这颗六面体是否均匀。

表 12-1 掷一颗六面体 300 次试验观测结果

点数 I	1	2	3	4	5	6
频数 O	43	49	56	45	66	41

(1) 数据录入有两种方式，分别参见数据文件 data12-01 和 data12-01a。数据文件 data12-01 是一种直接录入原始数据的方式，只有一个变量，在以下应用中可直接使用，但数据录入量较大。因此，

建议使用数据文件 data12-01a 的方式录入数据资料，则在进行下述操作方法前，应先用【数据】菜单中的【加权个案】过程将频数变量 Frequency 定义为权重变量。对变量“六面体[lmt]”进行加权处理。操作见第 2.4.1 节中的介绍。在本章的以后各例中，一律简称为“加权处理”。经加权处理后的变量，与数据文件 data12-01 方式录入的同名变量在后续的统计分析中是等价的。

(2) 操作方法。

- ① 读取数据文件 data12-01a。
- ② 按【分析→非参数检验→旧对话框→卡方】顺序打开【卡方检验】对话框，见图 12-3。
- ③ 选择“六面体[lmt]”变量进入【检验变量列表】框。
- ④ 由于这是一个均匀分布检验，故直接使用系统默认值，单击【确定】按钮，执行运算。

(3) 输出结果，见表 12-2。

表 12-2 六面体均匀性卡方检验结果

六面体				检验统计量	
	观察数	期望数	残差		六面体
1	43	50.0	-7.0	卡方	8.960 <sup>a</sup>
2	49	50.0	-1.0	df	5
3	56	50.0	6.0	渐近显著性	.111
4	45	50.0	-5.0	a.0个单元(0.0%)具有小于5的期望频率。单元最小期望频率为50.0。	
5	66	50.0	16.0		
6	41	50.0	-9.0		
总数	300				

说明：在本例中，原假设为：“这颗六面体是均匀分布的”。左表的第二列为实际观察值出



12.2 二项分布检验

二项分布检验 (Binomial Test) 是一种用来检验在给定的落入二项式中第一项概率值的前提下，数据来自二项分布的无效假设的方法。

12.2.1 二项分布检验的概念与操作

1. 二项分布检验的基本概念

如果随机变量  $X$  的分布如下：

$$P\{X = k\} = C_n^k p^k q^{n-k} \quad (k = 0, 1, 2, \dots, n) \quad (0 < p < 1, q = 1 - p)$$

则称  $X$  服从二项分布，或记为  $X \sim B(n, p)$ 。

在原假设  $H_0: p = p_0$  时，双侧精确检验的概率为

$$2 \left[ \sum_{i=0}^m \binom{N}{i} p^{*i} (1 - p^*)^{N-i} \right] - \binom{N}{m} p^{*m} (1 - p^*)^{N-m}$$

式中， $N = n_1 + n_2$ ， $n_1$  为类别 1 中观察值的数量， $n_2$  为类别 2 中观察值的数量，如果  $m = n_1$ ，则  $p^* = p$ ，否则， $p^* = 1 - p$ ； $p$  为检验的概率，而  $m = \min(n_1, n_2)$ 。

当  $p < \alpha$  时，拒绝原假设。 $\alpha$  一般取 0.05。

检验的变量应为数值型二分变量。如果是字符型变量，则先要使用【转换】菜单中的【自动重新编码】过程将其转换成数值型变量。【二分变量】是只能取两个值的变量，一般用 0 和 1 表示。在数据集中遇到的第一个值定义第一个组，其他值定义第二个组。如果变量不是二分变量，则必须指定分割点。利用分割点将小于或等于分割点的值的个案分到第一个组，并将其余个案分到第二个组。



图 12-6 【二项式检验】主对话框

在二项分布检验中，同样假定数据来自于随机样本。

2. 基本操作

(1) 按【分析→非参数检验→旧对话框→二项式】顺序打开如图 12-6 所示的对话框。

(2) 从变量列表选择一个或多个需要进行检验的变量，移到【检验变量列表】框中。

(3) 在【定义二分法】栏中定义二分值。

- ① 【从数据中获取】。适用于指定的变量只有两个有效值，为系统默认选项，无缺失值。
- ② 【割点】。如果指定的变量超过两个值，在参数框中输入分界点值，比分界点值小的形成第一项，否则构成第二项。
- (4) 在【检验比例】框中指定检验概率值。系统默认的检验概率是 0.5，这意味着要检验的二项是服从均匀分布的。如果落入每一项中的个体的期望比率不等，换言之，如果所要检验的二项不是等概分布，那么在参数框中输入与第一项所对应的概率期望值。

(5) 【精确】检验的选择、输出结果形式及缺失值处理方式，具体操作参见第 12.1.2 节相关内容。

(6) 单击【确定】按钮，执行命令；或单击【粘贴】按钮，在【语法】窗口生成命令语句。

12.2.2 二项分布检验分析实例

【例 3】 掷一枚球类比赛用的挑边器 31 次，出现 A 面、B 面在上的结果见表 12-5。取变量名为“tbh”，用数字型数据 1 代表“A”，用数字型数据 2 代表“B”，依次在数据库中输入数据。现检验这枚挑边器是否均匀。

表 12-5 掷一枚球类比赛用挑边器 31 次试验观测结果

次	1	2	3	4	5	6	7	8	9	10	11	12	13	13	15	16
面	A	B	A	B	B	A	A	A	B	B	A	B	B	A	A	A
次	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
面	B	A	B	B	A	B	B	A	B	A	B	B	A	B	A	

(1) 数据录入方式有两种，同【例 1】，分别存放在数据文件 data12-03a 和 data12-03 中，并对 data12-03a 中的 tbh 仿照【例 1】做法，进行了加权处理。

- (2) 操作方法。
- ① 读取数据文件 data12-03a 或 data12-03。
  - ② 按【分析→非参数检验→旧对话框→二项式】顺序打开如图 12-6 所示的对话框。
  - ③ 选择 tbh 变量进入【检验变量列表】框中。
  - ④ 由于所要检验的是这枚挑边器是否均匀，因此，这是一个均匀分布检验，故直接使用系统默认值。

⑤ 单击【确定】按钮，提交运算。

(3) 输出结果见表 12-6。

本例的原假设为：挑边器服从第一项的概率值为 0.5 的二项概率分布。在二项分布检验表中第二列列出的是分类编码，第三列为观察频数，第四列为各类的观察概率，第五列为检验概率，最后一列是在原假设为真的前提下，出现目前观察值及其更加极端值的概率。因  $p = 1.00 > 0.05$ ，表明目前证据不支持推翻原假设，故可认为这枚挑边器是均匀的。

表 12-6 挑边器均匀性二项分布检验结果

二项式检验					
	类别	N	观察比例	检验比例	精确显著性 (双侧)
挑边器	组 1	15	.48	.50	1.000
	组 2	16	.52		
	总数	31	1.00		

12.3 游程检验

12.3.1 游程检验的基本概念

一个游程就是某序列中位于一种符号之前或之后的另一种符号持续的最大主序列，或者说，一个游程是指某序列中同类元素的一个持续的最大主集。

例如，做一个掷硬币试验，以概率  $P$  得正面，以概率  $1-P$  得反面，用数字“0”记正面，用数字“1”记反面。不太可能出现多个 0 或多个 1 连续地连在一起，也不太可能 0 和 1 交替频繁地出现。假如做这样的试验 30 次，得到如下试验记录：

000011100000110000011111100000

如果称连在一起的 0 或连在一起的 1 为一个游程，则上面的例子中有 4 个 0 游程和 3 个 1 游程，共 7 个游程( $R=7$ )。记 0 出现的次数为  $n$ ，记 1 出现的次数为  $m$ ，则总的试验次数  $N=n+m$ 。显然，出现 0 或 1 的次数的多少同概率  $P$  有关，但在已知  $n$  和  $m$  时，游程数  $R$  的条件分布就同  $P$  无关了。

游程检验(Runs Test)就是根据游程数所作的两分变量的随机性检验。

其原假设为：两分变量有随机性。在原假设成立的前提下，当样本容量很大，即当  $m/n \rightarrow \gamma$  时

$$Z = \frac{R - \frac{2m}{1+\gamma}}{\sqrt{\frac{4\gamma m}{(1+\gamma)^3}}} \rightarrow N(0,1)$$

在给定显著性水平  $\alpha$  后，可用下面的近似公式得到临界值

$$c_1 = \frac{2mn}{m+n} \left[ 1 + \frac{\frac{Z_{\alpha/2}}{2}}{\sqrt{m+n}} \right] \qquad c_2 = \frac{2mn}{m+n} \left[ 1 - \frac{\frac{Z_{\alpha/2}}{2}}{\sqrt{m+n}} \right]$$

当计算得到的统计量的  $p$  值小于事先给定的显著性水平  $\alpha$  时，拒绝原假设。 $\alpha$  一般取 0.05。

游程检验可用来检验样本的随机性，这对于统计推断是很重要的。游程检验也可用来检验任何序列的随机性，而不管这个序列是怎样产生的。此外，它还可用来判断两个总体的分布是否相同，从而检验出它们的位置中心有无显著差异。

在具体的实际问题中，并不是所有的数据对都是以 0 或 1 的二元形式来表现的。例如，在遇到连续型的计量资料时，可先找出中位数，然后所有的原始数据与中位数来比较，大于中位数的计为 1，小于中位数的计为 0，这样可把计量资料变成一组 0、1 系列，就可按二元变量的随机性方法来做检验了。

如果样本来自的两总体的分布形态存在较大差距，则计算出的游程数会相对比较小。如果游程数比较大，则应是由于两样本数据充分混合的结果，它们的分布应该不存在显著差异。

用于游程检验的变量必须是数值型变量。如果是字符型变量，则先要使用【转换】菜单中的【自动重新编码】过程将其转换成数值型变量。此外，要求数据资料来自连续概率分布的样本。



图 12-7 【游程检验】主对话框

### 12.3.2 游程检验过程

(1) 按【分析→非参数检验→旧对话框→游程】顺序，打开【游程检验】对话框，见图 12-7。

(2) 从左侧变量列表选择一个或多个需要进行检验的变量，移到【检验变量列表】框中。

(3) 在【割点】栏内确定划分两类的分割点。在该框中提供了用来定义两类分割点的方法。变量值小于分割点的个体形成第一类，其他个体形成第二类。可选的分割点有【中位数】、【众数】、【均值】。还可以

选择【设定】复选项，将自定义分割点输入到后面的框中。

(4) 【精确】检验的选择、输出结果形式及缺失值处理方式，具体操作参见第 12.1.2 节相关内容。

(5) 单击【确定】按钮，执行命令。

12.3.3 游程检验分析实例

【例 4】 掷硬币 20 次得到的试验数据见表 12-7。其中，1 表示数字面朝上，0 表示画面朝上。建立变量 records，标签为“记录”，将表 12-7 中的数据录入到 SPSS 中，存在数据文件 data12-04 中。检验掷硬币试验是否是随机的。

1) 操作步骤

(1) 假设掷硬币的结果是随机的。

(2) 读取数据文件 data12-04。

(3) 按【分析→非参数检验→旧对话框→游程】顺序打开【游程检验】对话框。

(4) 选择 records 变量送入【检验变量列表】框中。

(5) 由于 1 出现 11 次，0 出现 9 次，中位数和众数都为 1，故不能选用【众数】和【中位数】作为划分两类的分割点。本例选用【均值】和【设定】（自定义期望值）。【割点】值应大于 0 且小于 1，本例输入 0.5。

(6) 单击【确定】按钮，提交运算。

2) 输出结果 (见表 12-8)

3) 输出结果解释

表 12-8 中，左边的游程检验表是以均值作为分界点的结果，计算结果平均数是 0.55；右边的游程检验表是割点 0.5 作为分界点的运行结果。

左边的游程检验表的第二行起依次为：检验值 0.55、小于检验值的样品数 9、大于等于检验值的样品数 11、总样品数 20、游程数量 12、Z 值 0.279 和双侧检验概率值 0.781。右边的游程检验表的第二行起依次为：检验值 0.50、总样品数 20、游程数量 12、Z 值为 0.279 和双侧检验概率值 0.781。两种检验有相同的  $p$  值。因  $p=0.781>0.05$ ，故不拒绝原假设，即掷硬币试验是随机的。

表 12-7 掷硬币 20 次的结果

1	1	0	1	0
0	0	1	1	0
1	0	1	1	1
0	0	1	1	0

表 12-8 掷硬币试验随机性检验结果

游程检验		游程检验	
	记录		记录
检验值 <sup>a</sup>	.5500	检验值 <sup>a</sup>	.5000
案例 < 检验值	9	案例总数	20
案例 ≥ 检验值	11	Runs 数	12
案例总数	20	Z	.279
Runs 数	12	渐近显著性(双侧)	.781
Z	.279	a. 用户指定的。	
渐近显著性(双侧)	.781		

12.4 一个样本的柯尔莫哥洛夫-斯米诺夫检验

12.4.1 一个样本的柯尔莫哥洛夫-斯米诺夫检验的基本概念

一个样本的柯尔莫哥洛夫-斯米诺夫检验 (One-Sample Kolmogorov-Smirnov Test) 简称单样本的 K-S 检验。它是用来检验样本来自正态分布、均匀分布或泊松分布总体的假设。这也是一

种拟合优度检验方法，它主要是运用某随机变量  $x$  的顺序样本来构造样本分布函数，使得能以一定的概率保证  $x$  的分布函数  $F(x)$  落在某个范围内。

K-S 双侧检验的原假设  $H_0$  为：对所有的  $x$  值  $F(x)=F(x_0)$  成立；备择假设为：至少有一个  $x$  值使  $F(x) \neq F(x_0)$  成立。

设  $S(x)$  表示一组数据的经验分布。定义一组随机样本  $x_1, x_2, \cdots, x_n$  的经验分布函数为阶梯函数

$$S(x)=\frac{x_i \leq x \text{ 个数}}{n}$$

它是小于  $x$  的值的比例，是总体分布  $F(x)$  的一个估计。检验统计量为

$$D=\sup_x|S(x)-F_0(x)|$$

$D$  的分布对一切连续分布  $f(x_0)$  在原假设下是一样的，所以它与分布无关。在实际运算中，由于  $s(x)$  是阶梯函数，只取离散值，所以考虑到跳跃问题，如果有  $n$  个观察值，则可用下面的统计量来代替上面的  $D$ ，即

$$D=\max_{1 \leq i \leq n}\left\{\max\left[|S(x_i)-F_0(x_i)|, |S(x_{i-1})-F_0(x_i)|\right]\right\}$$

当  $n \rightarrow \infty$  时，大样本的渐近公式为

$$P\left(\sqrt{n}D_n < x\right) \rightarrow K(x)$$

其分布函数的表达式为

$$K(x)=\begin{cases} 0 & x < 0 \\ \sum_{j=-\infty}^{\infty} (-1)^j \exp\left(-2j^2x^2\right) & x > 0 \end{cases}$$

当  $p < \alpha$  时，拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

需要注意的是，用本程序进行检验时，由于已知分布的参数是要用样本中计算得到的统计量来估计的，所以用渐近分布进行检验时，需要大样本，即  $n \geq 100$ 。当样本含量较少时，应选用精确检验或 Monte Carlo 检验。而当分布的参数已知时，则可以使用新版单样本中的 K-S 检验。它需先对已知参数进行设定，再进行检验，具体参见本章【例 10】中的相关作法。

12.4.2 柯尔莫哥洛夫-斯米诺夫检验过程

- (1) 按【分析→非参数检验→旧对话框→1-样本 K-S】顺序打开如图 12-8 所示对话框。
- (2) 从左侧变量表中选择一个或多个需进行检验的变量，移到【检验变量列表】框中。
- (3) 确定要检验的分布。【检验分布】框中提供了所要检验的分布，分别有【正态分布】(软件汉化为【常规】，不妥)、【均匀分布】(软件汉化为【相等】，不妥)、【泊松】分布、【指数分布】，系统默认检验正态分布。



图 12-8 【单样本 Kolmogorov-Smirnov 检验】主对话框



若要对指定参数的分布进行检验，则需要借助于 SPSS 命令语句。在 SPSS 命令语句中读者可以指定正态分布的平均数和标准差，为泊松分布指定平均数，为均匀分布指定最大值和最小值。

- (4) 精确检验的选择、输出结果形式及缺失值处理方式的操作参见第 12.1.2 节。
- (5) 单击【确定】按钮，执行命令。

12.4.3 柯尔莫哥洛夫-斯米诺夫检验分析实例

【例 5】数据文件 data12-05 是卢瑟福与盖革做的一个著名的试验的记录。他们观察由某块放射物质放出的，在 7.5 s 的时间间隔里到达某计数器的  $\alpha$  粒子数，共观察了 2608 次。表 12-9 中变量 zd 记录的是 7.5 s 到达计数器的粒子数，变量 fi 是每个 zd 值出现的次数。检验这种分布规律是否服从泊松分布。

表 12-9 质点试验数据

zd	0	1	2	3	4	5	6	7	8	9	10
fi	57	203	383	525	532	408	273	139	45	27	16

- (1) 数据文件 data12-05a 按表中数据录入，定义 fi 为加权变量。
- (2) 假设 7.5 s 内到达计数器的粒子数服从泊松分布。
- (3) 操作步骤。
  - ① 按【分析→非参数检验→旧对话框→1-样本 K-S】顺序单击菜单项，打开图 12-8 所示对话框。
  - ② 选择 zd 变量进入【检验变量列表】框。
  - ③ 在【检验分布】栏中选中【泊松】，对是否服从泊松分布进行检验。
  - ④ 单击【确定】按钮，提交运算。
- (4) 输出结果见表 12-10，给出了试验数据的泊松分布参数的计算结果。

表 12-10 Poisson 检验结果  
单样本 Kolmogorov-Smirnov 检验

		质点数
N		2608
a,b	均值	3.87
最极端差别	绝对值	.012
	正	.010
	负	-.012
Kolmogorov-Smirnov Z		.611
渐近显著性(双侧)		.850

检验分布为 Poisson 分布。  
根据数据计算得到。

12.5 两个独立样本检验

12.5.1 两个独立样本检验的用途与基本操作

两个独立样本均服从正态分布时比较均值使用 T 检验。但有时样本所隶属总体的分布类型可能不明或是非正态的，但还是想知道在这种情况下两个独立样本间是否具有相同的分布，两个独立样本检验(Two Independent Samples Test)就是用来处理此类问题的一种有效方法。执行本过程要求的数据文件结构与进行独立样本 T 检验的数据结构一样。

检验中用到的变量必须是可以排序的数值型变量。

- (1) 按【分析→非参数检验→旧对话框→两个独立样本】顺序单击菜单项，打开如图 12-9 所示的对话框。



图 12-9 【两个独立样本检验】主对话框

(2) 指定检验变量。从变量列表中选择要进行检验的一个或多个变量，移到【检验变量列表】框中。

(3) 指定分组变量名。从左面变量列表中指定用来分组的变量，并移到【分组变量】框中，单击【定义组】按钮，打开如图 12-10 所示的对话框，在两个编辑栏中输入分组值。

(4) 确定用来进行检验的方法。在【检验类型】框中，提供了可供用来检验的 4 种方法，检验两个独立样本(组)是否来自同一个总体。

①【Mann-Whitney U】。是非常流行的两个



图 12-10 【两独立样本：定义组】对话框

独立样本检验，等同于两组的 Wilcoxon 秩和检验和 Kruskal-Wallis 检验。此方法检验两个样本的总体在位置上是相等的。来自两组的观察被合并和赋予秩，有结的样品被赋予平均秩。结的数量相对于观察的总数量应该很少。如果总体在位置上是同样的，则秩应该是随机地混合在两个样本里的。换句话说，如果两个样本含量相同，则秩和也相同；如果两个样本含量不相同，则两个样本的平均秩相同。计算第一组得分领先第二组得分的次数和第二组得分领先第一组得分的次数。Mann-Whitney U 统计这两个数量的较少者。还显示较小样本秩和的 Wilcoxon 秩和  $W$  统计量。如果两个样本有相同的观察的数量，则  $W$  是在【定义组】对话框中首先指定的组[即【组 1】(Group1)组]的秩和。

所要作的原假设为  $H_0:F(x)=G(x)$ ，在原假设为真时，若  $\min\{m,n\} \rightarrow \infty$ ，且  $m/N \rightarrow \lambda \in (0,1)$ ， $\lambda$  是一个常数，则威尔科克森(Wilcoxon)秩和统计量  $W_y$  的概率分布和累积概率分布分别为

$$P(W_y = d) = \frac{t_{m,n}(d)}{\binom{N}{n}}$$
$$P(W_y \leq d) = \frac{\sum_{i \leq d} t_{m,n}(i)}{\binom{N}{n}} \quad d = n(n+1)/2, \dots, n(n+1)/2 + mn$$

式中， $t_{m,n}(d)$  表示从  $1, 2, \dots, N = m + n$  这  $N$  个数中任取  $n$  个数，其和恰为  $d$  的取法的种数。 $W_y$  的渐近正态性简记为

$$W_y \sim N[n(N+1)/2, mn(N+1)/12]$$

故

$$U = \frac{W_y - n(N+1)/2}{\sqrt{mn(N+1)/12}} \xrightarrow{L} N(0,1)$$

当观察值中有相等的值，即有结时，需通过对相等的观察值取平均秩，来对威尔科克森(Wilcoxon)秩和统计量  $W_y$  修正，此时， $W_y$  的渐近正态性为

$$W_y \sim N \left\{ n(N+1)/2, mn(N+1)/12 - nm \sum_{i=1}^g (t_i^3 - t_i) / [12N(N-1)] \right\}$$

式中,  $t_i$  为结的长度,  $i=1, 2, \dots, g$ 。

曼—惠特尼 (Mann-Whitney U) 统计量为

$$W_{xy} = W_y - n(n+1)/2, \quad W_{yx} = W_x - m(m+1)/2$$

它与威尔科克森 (Wilcoxon) 秩和统计量  $W_y$  只相差一个常数  $n(n+1)/2$ 。

同样在满足上述条件下,  $W_{xy}$  的概率分布和累积概率分别为

$$P(W_{xy} = d) = P(W_y = d + n(n+1)/2) = \frac{t_{m,n}[d + n(n+1)/2]}{\binom{N}{n}}$$

$$P(W_{xy} \leq d) = P(W_y \leq d + n(n+1)/2) = \frac{\sum_{i \leq d} t_{m,n}[i + n(n+1)/2]}{\binom{N}{n}} \quad d = 0, 1, \dots, mn$$

$W_{xy}$  的渐近正态性简记为

$$W_{xy} \sim N[mn/2, mn(N+1)/12]$$

$$U = \frac{W_{xy} - mn/2}{\sqrt{mn(N+1)/12}} \xrightarrow{L} N(0, 1), m \quad n \rightarrow \infty$$

同样, 在有结取平均秩时, 需对其方差作修正, 此时,  $W_{xy}$  的渐近正态性简记为

$$W_{xy} \sim N \left\{ mn/2, mn(N+1)/12 - nm \sum_{i=1}^g (t_i^3 - t_i) / [12N(N-1)] \right\}$$

当  $P < \alpha$  时, 拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

②【Kolmogorov-Smirnov Z】。是更普通的探测两者位置和分布形状上差异的检验。该检验是建立在两个样本的累积分布函数之间的最大绝对差异的基础上的。当这个差异显著地大时, 两个分布被认为是存在差异的。

假定从相互独立的连续型随机变量总体  $F_1(x)$  和  $F_2(x)$  中分别抽取样本  $x_1, x_2, \dots, x_{n_1}$  和样本  $y_1, y_2, \dots, y_{n_2}$ ,  $\hat{F}_1(x)$  和  $\hat{F}_2(x)$  分别是两个样本的对应累积经验分布函数。为计算经验分布函数及其差异, 首先将两组数据分别按由小到大的顺序排列成  $X_{[1]} \sim X_{[n]}$ , 第  $i$  组的经验累积函数用下式计算:

$$\hat{F}_i(X) = \begin{cases} 0 & -\infty < X < X_{[1]} \\ j/n_i & X_{[j]} \leq X < X_{[j+1]} \\ 1 & X_{[n]} \leq X < \infty \end{cases}$$

对两组中所有的  $X_j$  值, 两组间的差异为

$$D_j = \hat{F}_1(X_j) - \hat{F}_2(X_j)。$$

式中,  $\hat{F}_1(X_j)$  是对应于较大样本含量组的累积概率函数。同时计算最大正值、负值和绝对值。

所要检验的原假设为

$$H_0 : F_1(x) = F_2(x)。$$

Kolmogorov-Smirnov 提出的统计量为

$$D_{n_1, n_2} = \sup_{-\infty < x < \infty} \left| \hat{F}_1(x) - \hat{F}_2(x) \right|$$

当  $H_0$  为真时,  $Z = \sqrt{n}D_{n_1, n_2}$  有极限分布, 其中,  $n = n_1n_2 / (n_1 + n_2)$ 。

当  $n \rightarrow \infty$  时, 大样本的渐近公式为

$$P(\sqrt{n}D_{n_1, n_2} < x) \rightarrow K(x)$$

其分布函数的表达式为

$$K(x) = \begin{cases} 0 & x < 0 \\ \sum_{j=-\infty}^{\infty} (-1)^j \exp(-2j^2x^2) & x > 0 \end{cases}$$

当  $P < \alpha$  时, 拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

③【Moses 极限反应】。假设试验变量影响某个方向上的一些被试对象和相反方向上的其他被试对象。它检验同控制组比较的极端反应。本检验关键是控制组的跨度, 以及当合并控制组时, 试验组里有多少极端值影响跨度。分组对话框里【组 1】值定义的是控制组, 来自两个组的观察值被合并成一组进行排序并赋秩。

控制组的跨度 (SPAN) 是组中的最大和最小值的秩之差加 1, 取整到最接近的整数。由于跨度范围很容易受偶然因素的影响, 所以两端 5% 的样品被自动地删除。

所要检验的原假设为  $H_0 : F_1(x) = F_2(x)$ 。

精确的单侧概率水平用下式计算:

$$P(\text{SPAN} \leq n_c - 2h + g) = \sum_{i=0}^g \left[ \binom{i + n_c - 2h - 2}{i} \binom{n_c + 2h + 1 - i}{n_c - i} \right] / \binom{n_c + n_e}{n_c}$$

式中,  $h = 0$ ;  $n_c$  是控制组中样品的数量;  $n_e$  是试验组中样品的数量。同样的公式在下一步中使用, 那里  $h \neq 0$ 。如果用户不指定,  $h$  采用  $0.05 n_c$  的整数部分或 1 中的较大者, 如果用户指定, 除非它小于 1, 否则使用用户指定的整数值。

当  $P < \alpha$  时, 拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

④【Wald-Wolfowitz 游程】。该检验是更普通的探测在两者位置和分布形状上差异的检验。

假定从相互独立的连续型随机变量总体  $F_1(x)$  和  $F_2(x)$  中分别抽取样本  $x_1, x_2, \dots, x_{n_1}$  和样本  $y_1, y_2, \dots, y_{n_2}$ , 合并两个样本的所有观察值并由小到大排列为升序。计算对应于有序数据中同组数字变化的次数。游程数 ( $R$ ) 等于同组数字变化的次数加 1。如果两组观察数据中包含结, 计算可能的游程最小数量和最大数量。

如果两个样本是来自相同的总体, 则两组应被始终随机地分散分布。

因此, 所作的原假设为  $H_0 : F_1(x) = F_2(x)$ 。

如果总的样本量  $n_1 + n_2$ , 小于或等于 30, 单侧显著性水平可由下式中精确地计算:

当  $R$  是偶数时

$$P(r \leq R) = \frac{2}{\binom{n_1 + n_2}{n_1}} \sum_{r=2}^R \binom{n_1 - 1}{r/2 - 1} \binom{n_2 - 1}{r/2 - 1}$$

当  $R$  是奇数时

$$P(r \leq R) = \frac{1}{\binom{n_1 + n_2}{n_1}} \sum_{r=2}^R \binom{n_1 - 1}{k - 1} \binom{n_2 - 1}{k - 2} \binom{n_1 - 1}{k - 2} \binom{n_2 - 1}{k - 1}$$

式中， $r = 2k - 1$ 。

- 对于样本量大于 30，使用正态渐近分布。参见游程检验。
- 当  $P < \alpha$  时，拒绝原假设。显著性水平  $\alpha$  一般取 0.05。
- 在这 4 方法中，至少应选择一种。系统默认选项为【Mann-Whitney U】法。
- (5) 选择【精确】检验、输出结果形式及缺失值处理方式的操作参见第 12.1.2 节相关内容。
- (6) 单击【确定】按钮，提交运算。

12.5.2 两个独立样本检验分析实例

【例 6】设有甲、乙两种安眠药，考虑比较它们的治疗效果，独立观察 20 名患者。10 人服甲药，另 10 人服乙药，睡眠延长的时数见表 12-11。检验这两种药物的疗效有无显著性的差异。

表 12-11 两种安眠药效果对比数据

服甲药者睡眠延长时数	1.9	0.8	1.1	0.1	0.1	4.4	5.5	1.6	4.6	3.4
服乙药者睡眠延长时数	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0

- 延长的睡眠时数的分布情况不明，因此用非参数检验的方法。
- 数据文件 data12-06 中，变量 ycss 为服药后睡眠时间延长的时数；变量 zb 值为试验组别。组别用 1 表示服乙药、2 表示服甲药。录入数据后按变量 ycss 值降序排列。
- (1) 假设两组药物对延长睡眠时间的无显著差异。实际就是检验这两个独立样本是否具有相同的分布。
- (2) 操作步骤。
- ① 按【分析→非参数检验→旧对话框→2 个独立样本】顺序打开相应对话框。
- ② 选择 ycss 变量进入【检验变量列表】框中。
- ③ 选择 zb 变量进入【分组变量】框中。单击【定义组】按钮，打开【两独立样本定义组】对话框，在【组 1】框中输入“1”，在【组 2】框中输入“2”。
- ④ 由于在录入数据后已对数据作了排序处理，故在【检验类型】框中可选择其中的任何一种方法。本例选择了全部 4 种方法。
- ⑤ 单击【确定】按钮，提交运算。输出结果见表 12-12～表 12-15。
- 因 4 种方法计算的  $p$  值除 Mann-Whitney U 检验外均大于 0.05，故可认为这两种药物的疗效无显著性差异。

表 12-12 Mann-Whitney U 检验结果

秩					检验统计量 <sup>a</sup>	
	zb	N	秩均值	秩和		ycss
ycss	1	10	7.75	77.50	Mann-Whitney U	22.500
	2	10	13.25	132.50	Wilcoxon W	77.500
	总数	20			Z	-2.095
					渐近显著性(双侧)	.036
					精确显著性[Z* (单侧显著性)]	.035 <sup>b</sup>
a. 分组变量: zb						
b. 没有对结进行修正。						

表 12-13 Moses 检验结果

频率		检验统计量 <sup>a,b</sup>		
ycss	zb	N	ycss	
1 (控制)		10	控制组观察跨度	17
2 (试验)		10	显著性 (单侧)	.291
总数		20	修整的控制组跨度	15
			显著性 (单侧)	.686
			从每个末端修整的离群者	1
a. Moses 检验				
b. 分组变量: zb				

表 12-14 两样本 K-S Z 检验结果

频率		检验统计量 <sup>a</sup>	
zb	N		ycss
1	10	最极端差别 绝对值	.500
2	10	正	.500
总数	20	负	.000
		Kolmogorov-Smirnov Z	1.118
		渐近显著性(双侧)	.164

a. 分组变量: zb

表 12-15 Wald-Wolfowitz 检验结果

频率		检验统计量 <sup>a,b</sup>			
ycss	N		Runs 数	Z	精确显著性 (单侧)
1	10	最小可能	6 <sup>c</sup>	-2.068	.019
2	10	最大可能	10 <sup>c</sup>	-.230	.414
总数	20				

a. Wald-Wolfowitz 检验  
b. 分组变量: zb  
c. 有 2 个涉及 7 个案例的组间结。

12.6 多个独立样本检验

12.6.1 多个独立样本检验的用途与操作

前面所提到的两个独立样本检验是多个独立样本检验中最基本的形式,要解决多个独立样本是否具有相同的分布的问题,需借助于多个独立样本检验方法。检验中用到的变量必须是可以排序的数值型变量。操作方法如下:

(1) 按【分析→非参数检验→旧对话框→K 个独立样本】顺序打开如图 12-11 所示的对话框。



图 12-11 【多个独立样本检验】主对话框

(2) 指定检验变量。

从左侧变量列表选择一个或多个需要进行检验的变量,移到【检验变量列表】框中。

(3) 指定分组变量值范围。

从变量列表中选择分组变量移到【分组变量】框中,单击【定义范围】按钮,在相应的对话框中,输入最小值和最大值来定义分组变量值范围。

(4) 在【检验类型】选项中确定用来进行检验的方法有 3 种。

①【Kruskal-Wallis H】检验。是 Mann-Whitney U 检验的扩展,类似单因素方差分析,探究分布位置上的差异。该方法假设从  $k$  个无序的总体中抽取样本,是系统默认的方法。它所检验的问题称为无方向检验问题。

所要检验的原假设为  $H_0: \theta_1 = \theta_2 = \dots = \theta_k$ , 即  $k$  个位置参数相等。

对于有结时,不校正的检验统计量为

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k R_i^2 / n_i - 3(N+1)$$

对于有结时,校正的统计量为

$$H' = \frac{H}{1 - \sum_{i=1}^m T_i / (N^3 - N)}$$

式中,  $N = \sum_{i=1}^k n_i$ ,  $n_i$  为第  $i$  组的观察值的数量;  $R_i$  为第  $i$  组的秩和;  $T_i$  为第  $i$  组的结的长度,  $T_i = t_i^3 - t_i$ ,  $m$  为结集的数量。

当原假设为真时,统计量  $H$  和  $H'$  渐近服从  $\chi^2(k-1)$  分布。故显著性水平基于自由度为  $k-1$  的  $\chi^2$  分布。

当  $P < \alpha$  时, 拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

②【中位数】检验。是很普通的检验, 但效率不高, 探究在位置和形状上的分布差异。该方法假设从  $k$  个无序的总体中抽取样本。它也是用来检验无方向问题的。

如果中位数未被使用者指定, 则它通过所有组合并数据排序后按下面公式计算确定:

$$M_d = \begin{cases} (X_{[N/2]} + X_{[N/2+1]}) / 2 & \text{如果 } N \text{ 是偶数} \\ X_{[(N+1)/2]} & \text{如果 } N \text{ 是奇数} \end{cases}$$

式中,  $X_{[N]}$  为最大值,  $X_{[1]}$  为最小值。

设共有  $k$  个组, 记第  $i$  个组的小于等于中位数的个数为  $O_{1i}$ , 大于中位数的个数为  $O_{2i}$ ,

$$R_1 = \sum_{i=1}^k O_{1i}, \quad R_2 = \sum_{i=1}^k O_{2i}, \quad n_i = O_{1i} + O_{2i}, \quad N = R_1 + R_2.$$

所要检验的原假设为  $H_0: M_{d_1} = M_{d_2} = \cdots = M_{d_k}$ 。

$\chi^2$  统计量用下式计算:

$$\chi^2 = \sum_{j=1}^k \sum_{i=1}^2 (O_{ij} - E_{ij})^2 / E_{ij}$$

$$\text{式中, } E_{ij} = \frac{R_i n_j}{N}.$$

在  $H_0$  为真时, 上式确定的  $\chi^2$  近似服从  $\chi^2(k-1)$  的卡方分布。

当  $P < \alpha$  时, 拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

③【Jonckheere-Terpstra】(乔卡契尔-特普斯特拉) 检验。当  $k$  个总体有序(升序或降序)时, 此检验方法非常有效。例如,  $k$  个总体可以描述  $k$  个增加的温度。检验的假设是不同的温度产生同样反应的分布, 备择假设: 温度升高反应剧烈。这里, 假设两个样本是有序的, 因此, 使用 Jonckheere-Terpstra 检验是最适当的。安装了精确检验(Exact Tests)时此检验才是可选用的。

乔卡契尔-特普斯特拉检验的基本思想如下。

设有  $k$  个连续型随机变量总体  $X_1, X_2, \cdots, X_k$ ,  $x_{i1}, x_{i2}, \cdots, x_{in_i}$  是来自第  $i$  个总体  $X_i$  的样本, 其样本量为  $n_i$ ,  $i=1, 2, \cdots, k$ , 则总的样本容量为  $N = \sum_{i=1}^k n_i$ 。所有  $N$  个样本单元都是相互独立的。

设第  $i$  个总体  $X_i$  的分布函数为  $F(x - \theta_i)$ ,  $i=1, 2, \cdots, k$ 。对于有方向性的检验问题, 所要检验的原假设  $H_0: \theta_1 = \theta_2 = \cdots = \theta_k$ ,  $H_1: \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$ , 且  $\theta_1 < \theta_k$ 。

Jonckheere-Terpstra 检验用  $J$  作为统计量,  $J = \sum_{1 \leq i < j \leq k} W_{ij}$ , 其中的  $W_{ij}$  就是 Mann-Whitney  $U$

统计量。  $W_{ij} = \# \{(x_{ir}, x_{js}): x_{ir} < x_{js}, r=1, 2, \cdots, n_i; s=1, 2, \cdots, n_j\}$ , “#” 表示计数。

可以证明, 在原假设为真时, 若  $\min\{n_1, \cdots, n_k\} \rightarrow \infty$ , 且对所有的  $i=1, 2, \cdots, k$ , 都有  $n_i / N \rightarrow \lambda_i \in (0, 1)$ , 则

$$J - T = \frac{J - E(J)}{\sqrt{D(J)}} = \frac{J - \frac{1}{4} \left( N^2 - \sum_{i=1}^k n_i^2 \right)}{\sqrt{\frac{1}{72} \left[ N^2 (2N+3) - \sum_{i=1}^k n_i^2 (2n_i+3) \right]}} \xrightarrow{L} N(0, 1)$$

当全部样本中结的个数为  $g$  时, 需对上式中的  $D(J)$  部分作如下修正:

$$D(J)=\frac{1}{72}\left[N^2(2N+3)-\sum_{i=1}^kn_i^2(2n_i+3)-\sum_{s=1}^gt_s(t_s-1)(2t_s+5)\right]+\frac{1}{36N(N-1)(N-2)}\left[\sum_{i=1}^kn_i(n_i-1)(n_i-2)\right]\left[\sum_{s=1}^gt_s(t_s-1)(t_s-2)\right]+\frac{1}{8N(N-1)}\left[\sum_{i=1}^kn_i(n_i-1)\right]\left[\sum_{s=1}^gt_s(t_s-1)\right]$$

当  $P < \alpha$  时, 拒绝原假设。显著性水平  $\alpha$  一般取 0.05。  
(5) 选择精确检验、输出结果形式及缺失值处理方式的操作参见第 12.1.2 节相关内容。

12.6.2 多个独立样本检验分析实例

【例 7】 某车间用 4 种不同的操作方法各做若干批试验, 试验中优等品率(%) 数据资料见表 12-16, 检验操作方法对产品的优等品率是否有显著影响。

表 12-16 4 种不同操作方法的优等品率试验数据

试验批号	操作方法 1	操作方法 2	操作方法 3	操作方法 4
1	12.1	18.3	13.7	7.3
2	14.8	49.6	25.1	1.9
3	15.3	10.1	47.0	5.8
4	11.4	35.6	16.3	10.1
5	10.8	26.2	30.4	9.4
6		8.9		

- (1) 数据在文件 data12-07 中, 变量 ydp1 为优等品率, ff 为操作方法。
- (2) 操作步骤。
- ① 按【分析→非参数检验→旧对话框→K 个独立样本】顺序打开主对话框。
- ② 选择 ydp1 变量进入【检验变量列表】框中。
- ③ 选择 ff 变量进入【分组变量】框中。单击【定义范围】按钮, 打开【多个独立样本检验: 定义范围】对话框, 在【最小值】框中输入“1”, 在【最大值】框中输入“4”。
- ④ 在【检验类型】框中选择 Kruskal-Wallis H 法。因每组观测值数量太少, 【中位数】法不适用, 故不选择。
- ⑤ 单击【确定】按钮, 提交运算。
- (3) 输出结果见表 12-17。检验结果包括两个表。

表 12-17 Kruskal-Wallis 检验结果<sup>a,b,c</sup>

ff	N	秩均值
yp1	5	10.40
2	6	13.75
3	5	15.80
4	5	3.50
总数	21	

	yp1
卡方	11.530
df	3
渐近显著性	.009

a Kruskal Wallis 检验  
b 分组变量: ff  
c 由于没有足够内存, 无法计算某些或所有精确显著性。

在 Kruskal-Wallis H 法中, 计算的  $p$  值约等于 0.009, 小于 0.05, 这表明, 拒绝 4 种操作方法有相同的产品优等率的原假设, 而去接受其相反的备择假设时, 犯错误的概率不足 0.009, 故可认为这 4 种不同的操作方法对产品优等品率是有显著影响的。



## 12.7 两个相关样本检验

### 12.7.1 两个相关样本检验的用途与操作

在实际的研究工作中，经常会遇到从同一个被测试对象上测试两个或多个观测值的情况，这样的数据间就不再是相互独立的了，而是彼此相关。在此种情况下，检验样本间是否具有相同的分布，要用两个相关样本检验。

检验中用到的变量必须是可以排序的数值型变量。

(1) 按【分析→非参数检验→旧对话框→两个相关样本】顺序打开如图 12-12 所示的对话框。

(2) 指定检验变量对。

从左面变量列表中同时选择两个待检验的变量，送到【检验对】框中，在【对】的【Variable 1】和【Variable 2】中依次出现所选择的两个变量名。如果相关的成对变量为多组，则重复上述操作。

(3) 确定检验方法。

本节里检验比较两个相关变量的分布。检验的方法根据数据类型确定。在【检验类型】框中，提供了 4 种方法。



图 12-12 【两个关联样本检验】主对话框

① 如果是连续型数据，则使用符号检验 (Sign test) 或威尔科克森符号等级检验 (Wilcoxon Signed-Rank test)。

- 【符号检验】。对所有样品计算两个变量值间的差值，并将差值分为正、负或结(相等)3 类。如果两个变量有类似的分布，则正、负数的数目上的差异应无显著不同。

符号检验的基本做法如下：

对于每个样品，计算差异  $D_i = X_i - Y_i$ ，合计正差异的数量  $S_p$  和负差异的数量  $S_n$ ，对于  $X_i = Y_i$  的样品，不算在正、负之列。在满足一定的条件下， $D_i$  独立同分布，因此，两个样本间有无差异的检验问题，就等价于对称中心  $\theta$  是否等于 0 的检验问题。

所要作的原假设为  $H_0$ ：对称中心  $\theta = 0$ 。

如果  $n_p + n_n \leq 25$ ，当  $p = 0.5$  和  $r = \min(n_p + n_n)$  时，在  $n_p + n_n$  次试验中， $r$  或少数“成功”事件的精确概率用下面的二项分布公式递推计算：

$$P(X \leq r) = \sum_{i=0}^r \binom{n_p + n_n}{i} (0.5)^{(n_p + n_n)}$$

如果  $n_p + n_n > 25$ ，则显著性水平根据正态近似值

$$Z_c = \frac{\max(n_p, n_n) - 0.5(n_p + n_n) - 0.5}{0.5\sqrt{n_p + n_n}} \xrightarrow{L} N(0,1),$$

计算得到。

当  $P < \alpha$  时，拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

- **【Wilcoxon】** (威尔科克森符号等级检验)。要求比较的两个变量分布形状相似，它不但考虑两个符号数目上的差异，而且还考虑成对样品数值之间差异幅度的因素。由于 Wilcoxon 符号秩检验纳入了有关数据的更多信息，因此它比符号检验更为强大。

威尔科克森 (Wilcoxon) 符号等级检验的基本做法如下：

对于每个样品，在排序前，先计算成对观察值之间的差异  $D_i = X_i - Y_i$  和  $|D_i|$ 。将所有非零的绝对差排列成升序并赋秩。在有结的情况下，对结点使用平均秩。计算对应于正差异的秩和  $S_p$  和负差异的秩和  $S_n$ 。正秩的均值为

$$\bar{X}_p = S_p / n_p$$

负秩的均值为

$$\bar{X}_n = S_n / n_n$$

式中， $n_p$  是有正差异的样品的数量； $n_n$  是有负差异的样品的数量。

所要作的原假设为  $H_0$ ：对称中心  $\theta = 0$ 。

在原假设为真时，大样本下，统计量

$$Z = \frac{\min(S_p, S_n) - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24 - \sum_{j=1}^l (t_j^3 - t_j)/48}} \xrightarrow{L} N(0,1)$$

式中， $n$  为非零差异样品的数量； $l$  为结的数量； $t_j$  为第  $j$  个结的长度。

当  $P < \alpha$  时，拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

② 如果是二分数据，则使用 **【McNemar】** 检验。每个被试对象的响应分别在指定事件发生的前、后被重复测定。McNemar 检验确定初始的响应率(事件前)是否等于最终响应率(事件后)。本检验对于在前后对比设计中检测由试验干预引起的响应变化很有用。

McNemar 检验的基本做法如下：

被研究的数据值限定为两个唯一响应类别。合计  $X_i < Y_i$  的样品数  $n_1$  或  $X_i > Y_i$  的样品数  $n_2$ 。

如果  $n_1 + n_2 \leq 25$ ，当  $P = 0.5$  和  $r = \min(n_1 + n_2)$  时，在  $n_1 + n_2$  次试验中， $r$  或少数“成功”事件的精确概率用下面的二项分布公式递推计算：

$$P(X \leq r) = \sum_{i=0}^r \binom{n_1 + n_2}{i} (0.5)^{(n_1 + n_2)}$$

双侧检验的概率水平用计算得到的值加倍获得。如果  $n_1 + n_2 > 25$ ，则使用连续性修正的  $\chi^2$  近似值：

$$\chi_c^2 = \frac{(|n_1 - n_2| - 1)^2}{n_1 + n_2} \sim \chi^2(1)$$

当  $P < \alpha$  时，拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

③ 如果是分类数据，则使用**【边际同质性】**检验 (Marginal Homogeneity test)。它是 McNemar 检验从二分响应向多重响应的扩展。它使用卡方分布检验试验干涉前、后设计中反应的变化。对于在前后对比设计中检测因试验干预所导致的响应变化很有用。边缘同质检验只有在安装了**【精确检验】**附件时，才有效。

(4) 【精确检验】方法的选择、输出结果形式及缺失值处理方式的选择操作见 12.1.2 节相关内容。

(5) 单击【确定】按钮，提交运算。

12.7.2 两个相关样本检验分析实例

【例 8】为研究长跑运动对增强普通高校学生的心功能效果，对某校 15 名男生进行测试，经过 5 个月的长跑锻炼后看其晨脉是否减少。锻炼前、后的晨脉数据见表 12-18。检验锻炼前、后的晨脉间有无显著性的差异。

表 12-18 长跑锻炼前、后的晨脉变化

锻炼前	70	76	56	63	63	56	58	60	65	65	75	66	56	59	70
锻炼后	48	54	60	64	48	55	54	45	51	48	56	48	64	50	54

- (1) 数据文件 data12-08 中，变量 dlq 为锻炼前的晨脉，变量 dlh 为锻炼后的晨脉。
- (2) 操作步骤。

① 按【分析→非参数检验→旧对话框→两个相关样本】顺序打开主对话框。

② 选择 dlq 和 dlh 变量进入【检验对】框中。

③ 由于晨脉数据为连续型数据，因此在【检验类型】框中选择【Wilcoxon】和【符号检验】。

④ 单击【确定】按钮，提交运算。
- (3) 输出结果见表 12-19 和表 12-20。

表 12-19 Wilcoxon Signed Ranks 检验结果

秩

	N	秩均值	秩和
pulse after - pulse befor 负秩	12 <sup>a</sup>	9.17	110.00
正秩	3 <sup>b</sup>	3.33	10.00
结	0 <sup>c</sup>		
总数	15		

a. pulse after < pulse befor

b. pulse after > pulse befor

c. pulse after = pulse befor

Test Statistics<sup>a</sup>

	晨练后脉搏 - 晨练前脉搏
Z	-2.842 <sup>a</sup>
Asymp. Sig. (2-tailed)	.004

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

在本检验过程中，锻炼后的晨脉数据(或其秩)减锻炼前的晨脉数据(或其秩)，若大于 0 则取正号，小于 0 取负号，等于 0 为结。两种检验方法中，得到的负号数量均为 12，正号数量均为 3，而结都为 0，因两种检验方法计算的  $p$  值均小于 0.05，表明现有证据足于支持拒绝锻炼前、后的晨脉间分布相同的原假设，故可认为锻炼前、后的晨脉间有显著性的差异。

表 12-20 Sign 检验结果

频率

		N
pulse after - pulse befor	负差分 <sup>a</sup>	12
	正差分 <sup>b</sup>	3
	结 <sup>c</sup>	0
	总数	15

a. pulse after < pulse befor

b. pulse after > pulse befor

c. pulse after = pulse befor

检验统计量<sup>a</sup>

	pulse after - pulse befor
精确显著性 (双侧)	.035 <sup>b</sup>

a. 符号检验

b. 已使用的二项式分布。

## 12.8 多个相关样本检验

### 12.8.1 多个相关样本检验的用途与操作

两个相关样本检验是多个相关样本检验最基本的形式,要解决多个相关样本间是否具有相同的分布的问题,使用多个相关样本检验(Test for Several Related Samples)方法。

检验中用到的变量必须是可排序的数值型变量。

- (1) 按【分析→非参数检验→旧对话框→K 个相关样本】顺序打开如图 12-13 所示的对话框。
- (2) 从源变量列表中选择需要进行检验的一个或多个变量,移到【检验变量】框中。
- (3) 在【检验类型】栏中,根据变量类型确定检验方法。

①【Friedman】是等同于一个样本重复测定设计或每单元一个观察值的双向方差分析的非参数检验。

其基本做法为:对  $N$  个样品中的每个样品,  $k$  个变量被排序并被从  $1 \sim k$  赋秩,在结上赋予平均秩。对  $k$  个变量中的每个变量,计算样品的秩和。用符号  $C_i$  表示,则每个变量的平均秩为  $R_i = C_i / N$ 。

所要作的无效假设为  $H_0:k$  个相关的变量来自同一个总体。

检验统计量为

$$\chi^2 = \frac{[12 / Nk(k+1)] \sum_{i=1}^k C_i^2 - 3N(k+1)}{1 - \sum T / [Nk(k^2 - 1)]}$$

式中,  $\sum T = \sum_{i=1}^N \sum_{l=1}^k (t^3 - t)$ ,  $t$  是变量结的长度。

在原假设为真时,上面的  $\chi^2 \sim \chi^2(k-1)$ 。

当  $P < \alpha$  时,拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

②【Kendall 的 W】是标准化的 Friedman 统计量,是调和系数。它是比率之间一致性的测度。每个样品是一个鉴定人或定价人,每个变量是一个条件或被鉴定的人。对每个变量计算秩和。Kendall's W 范围在 0(不同意)~1(同意)之间。

协(调)和系数  $W$  用下式计算

$$W = \frac{F}{N(k-1)} \frac{N^2 k(k^2 - 1) / 12}{N^2 k(k^2 - 1) / 12 - N \sum T / 12}$$

式中,  $F$  是 Friedman 检验中的  $\chi^2$  统计量;  $\sum T = \sum_{i=1}^N \sum_{l=1}^k (t^3 - t)$ ,  $t$  是变量结的长度;  $N$ 、 $k$  和

$l$  的含义同 Friedman 检验。

它所检验的无效假设为  $H_0:\theta_1 = \theta_2 = \dots = \theta_k$ , 备择假设为  $H_1:\theta_1, \theta_2, \dots, \theta_k$  不全相等。

在原假设为真时,  $\chi^2 = N(k-1)W \sim \chi^2(k-1)$ 。

当  $P < \alpha$  (显著性水平  $\alpha$  一般取 0.05)时,拒绝原假设。则认为各  $\theta_i$  之间有一个顺序关系,



图 12-13 【多个关联样本检验】主对话框

也即这  $k$  个观察值有这样的趋势:  $x_{1j} \leq x_{2j} \leq \cdots \leq x_{kj}, x_{1j} < x_{kj}$ 。这说明任意第  $j$  个区组内的  $k$  个观察值都有这样的趋势, 所以在  $b$  个区组中一致性趋于成立。

③【Cochran 的 Q】同 Friedman 检验是相同的, 适用于所有的应答是二值的情况。它是将 McNemar 检验扩展到  $k$  个样本的一种检验方法。Cochran's Q 检验假设几个相关的两分变量有相同的均数。变量是在同一个个体或在配对个体上测定的。

其基本做法为: 对  $N$  个样品中的每一个样品, 在  $k$  个指定的两分变量上取值, 两分变量上第一个取到的值被当作“成功”处理, 对每个样品合计“成功”变量的数量。样品  $i$  “成功”的数量用  $R_i$  标记, 变量  $l$  的总的“成功”的数量用  $C_l$  标记。

Cochran's Q 检验所作的原假设为  $H_0$ : 几个相关的两分变量有相同的均数。

Cochran's Q 用下式计算:

$$Q = \frac{(k-1) \left[ k \sum_{l=1}^k C_l^2 - \left( \sum_{l=1}^k C_l \right)^2 \right]}{k \sum_{l=1}^k C_l - \sum_{i=1}^N R_i^2}$$

当原假设为真时,  $Q \sim \chi^2(k-1)$ 。

当  $P < \alpha$  时, 拒绝原假设。显著性水平  $\alpha$  一般取 0.05。

(4) 选择【精确】检验方法、输出结果形式及缺失值处理方式的操作参见第 12.1.2 节相关内容。

(5) 单击【确定】按钮, 提交系统执行。

12.8.2 多个相关样本检验分析实例

【例 9】某商店想了解顾客对几种不同款式衬衣的喜爱程度。某日询问了 9 名顾客, 请他们对 3 种款式的衬衣按喜爱程度排次序(最喜爱的给秩 1, 其次的给秩 2, 再次的给秩 3), 结果见表 12-21。检验顾客对 3 种款式衬衣的喜爱程度是否相同。对应数据文件为 data12-09。变量  $a$ 、 $b$ 、 $c$  的数据分别是顾客对款式 A、B、C 衬衣的喜爱程度。

表 12-21 顾客对不同款式衬衣喜爱程度

顾客号	1	2	3	4	5	6	7	8	9
款式 A	2	2	2	1	3	1	2	1	1
款式 B	3	3	3	3	2	2	3	3	3
款式 C	1	1	1	2	1	3	1	2	2

(1) 假设顾客对 3 种款式衬衣的喜爱程度无显著差异。

(2) 按【分析→非参数检验→旧对话框→K 个相关样本】顺序打开相应的对话框。选择  $a$ 、 $b$ 、 $c$  变量进入检验变量框中。在【检验类型】框中选择【Friedman】方法和【Kendall 的 W】方法。单击【确定】按钮, 提交运算。

(3) 输出结果见表 12-22 和表 12-23。

表 12-22 Friedman 检验结果

秩		检验统计量 <sup>a</sup>	
	秩均值	N	9
score of A	1.67	卡方	8.222
score of B	2.78	df	2
score of C	1.56	渐进显著性	.016

a. Friedman 检验

表 12-23 Kendall's W 检验结果

秩		检验统计量	
	秩均值	N	9
score of A	1.67	Kendall W <sup>a</sup>	.457
score of B	2.78	卡方	8.222
score of C	1.56	df	2
		渐进显著性	.016

a. Kendall 协同系数

因两种检验方法计算的  $p$  值均等于 0.016，小于 0.05，故可认为顾客对 3 种款式的衬衣的喜爱程度是不相同的。

## 12.9 新版非参数假设检验的界面及其使用方法

这里所谓的“新版”是针对“旧对话框”中的界面而言的，而非 SPSS 版本的新旧。新版的非参数检验以按单样本检验（分布的一致性检验）、独立样本检验和相关样本检验为主线重新整理归类方法、设计界面后而成。因此，在旧对话框过程中能处理的问题，同样可在新版的某个过程的相应方法中得以实现。

旧版对话框中的卡方检验、二项式（分布）检验、游程检验、1-样本 K-S 检验合并到新版的单样本检验过程，两个独立样本检验、K 个独立样本检验合并到新版的独立样本检验过程，两个相关样本检验、K 个相关样本检验合并到新版的相关样本检验过程。

新版除对旧对话框中的方法进行重新归类整理外，在功能方面也有一些更新。具体体现在：在所要分析的数据文件中，如果对变量的角色功能已根据需要进行了预定义，那么，新版已具备自动检查变量角色属性，可据此设定检验变量与所要比较的分类变量及自动选择检验方法的功能。这是新版最大的改进之处。此外，在关于一个样本与已知总体参数的分布一致性检验中，新版已不需要像旧版那样，用编写 SPSS 命令语句的方式来实现，而只需直接在选择检验方法对话框的相应参数框中设定分布参数即可。在输出内容方面，新版在输出中增加了检验模型浏览器功能，使得输出中图表兼备，直观明了。

需要指出的是：在新版中所谓的“字段”就是在前面章节中提到的“变量”的另一种称谓。它是计算机语言中对变量的习惯称谓。

### 12.9.1 单样本检验

#### 12.9.1.1 单样本检验的用途

单样本非参数检验使用一个或多个非参数检验来识别单个变量与给定分布之间的差别。它不需要假定检验变量的数据资料呈正态分布。



图 12-14 【单样本非参数检验】对话框【目标】选项卡

#### 12.9.1.2 单样本检验的操作

按【分析→非参数检验→单样本】顺序打开如图 12-14 所示的【单样本非参数检验】的【目标】选项卡。

##### 1. 在【目标】选项卡中设定检验目标

在【您的目标是什么？】栏中，可快速指定常用的不同检验设置。每个目标对应【设置】选项卡上的一个默认配置。如有需要可以自定义配置。

(1) 【自动比较观察数据和假设数据】。这是系统默认选项，选择该选项，单样本检验过程自动对二分变量使用二项分布检验，对所有其他分类变量使用卡方检验，对连续型变量使用 Kolmogorov-Smirnov 检验。

值得注意的是,新版具有自动识别功能,但它是建立在对变量的测度类型进行明确定义的基础上的。如果在检验之前已经对变量的测度类型进行了准确定义,则在新版中的该选项能省去在选择检验方法上的许多烦恼,否则,在旧对话框中能实现的检验,在新版中就不能完成。

(2) 【检验随机序列】。单样本检验过程使用游程检验来检验观察到的数据值序列的随机性。

(3) 【自定义分析】。如果希望手动修改“设置”选项卡上的检验设置,应选择该选项。但如果随后在“设置”选项卡上更改了与当前选定目标不一致的选项,系统则会自动选择该设置。例如,在【目标】选项卡中选择【自定义分析】,但在后面提到的【设置】选项卡中,未作设置,而直接单击了【运行】按钮,那么,这相当于采用了系统默认的【自动比较观察数据和假设数据】选项,而没有选择一个真正想选的检验方法。

## 2. 在【字段】选项卡中选择要进行检验的变量

单击【字段】按钮,打开【字段】选项卡,见图 12-15。

在【字段】选项卡中对需要进行检验的字段进行指定。至少要选择一個需要进行检验的变量。

(1) 【使用预定义角色】。此选项使用现有的变量信息。

【字段】选项卡支持可用于预先选择分析变量的预定义角色。

在 SPSS 中,变量所扮演的角色一般分为输入(如预测变量、自变量)、输出或目标(如因变量)、同时用作输入和输出、没有角色分配、分区及拆分等。参见第 2 章相关内容。

当打开【字段】选项卡时,满足角色要求的变量将自动显示在目标列表中(即【字段】框和【检验字段】框中)。默认情况下,为所有变量分配输入角色。角色分配只影响支持角色分配的对话框,对命令语法没有影响。

所有预定义角色为“输入”、“目标”或“两者”(变量将同时用作输入和输出)的字段将用作检验字段(变量)。

(2) 【使用定制(自定义)字段分配】。使用本选项,可以将【字段】框中列出的所要进行检验的变量移入【检验字段】框中进行指定。至少需选择一个字段(变量)。它可以避免【使用预定义角色】时,极可能出现将不需要检验的字段都选作检验变量的情况。

根据需要,用户可对出现在【字段】框中的变量列表按一定方式进行排序。在【排序】下拉列表(见图 12-16)中,共有 3 种排序方式:

① 【无】。系统默认方式。在该方式下,是按在数据集中变量名出现的先后顺序进行排列的。



图 12-15 【字段】选项卡

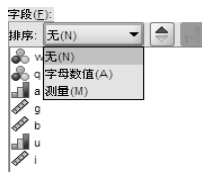





图 12-16 字段排序方式

②【字母数值】。按字母的 ASCII 码和数值大小顺序进行有序排列。

③【测量】(测度标准)。按测度标准即名义、有序、尺度顺序进行排序。在同一个测度标准中,则按它们在数据集中变量名出现的先后顺序进行排列。

(3) 单击【字段】列表框下面的全部(A)按钮,选中【字段】列表框中的所有变量;单击按钮,则选中【字段】列表框中的所有名义变量;单击按钮,选中【字段】列表框中的所有有序变量;单击按钮,选中【字段】列表框中的所有尺度变量。

### 3. 在【设置】选项卡中设置检验方法及其选项

单击【设置】按钮,打开【设置】选项卡,见图 12-17,可对在【字段】选项卡中指定所选【字段】需要执行的检验及其选项进行设置。

在【选择项目】栏中,需对【选择检验】、【检验选项】和【用户缺失值】3 个方面分别进行设置。选择一个项目打开相应的对话框。

#### (1) 选择检验。

在系统默认情况下,【设置】选项卡界面处于选择检验的状态。此时,可对检验类型进行设定。它有两个选择,系统默认选择为【根据数据自动选择检验】。

一般情况下,只有在单样本非参数检验的【目标】选项卡中,选择了【自动比较观察数据和假设数据】选项时,才能够对应选择此项。如果选择此项,则意味着将对二分变量(字段)进行二项分布检验,而对所有其他分类的变量(字段)应用卡方检验,对连续型变量(字段)应用 Kolmogorov-Smirnov 检验。

另一个选择为【自定义检验】。在【自定义检验】中,有【二项分布检验】、【卡方检验】、【Kolmogorov-Smirnov 检验】、【Wilcoxon 符号秩检验】及【游程检验】可供选择。

各种检验方法及其适用条件已在旧对话框的相关章节中介绍,不再赘述。

下面对各种方法涉及的界面及其选项作相应的介绍。

①【二项分布检验】。选择【比较观察二分类可能性和假设二分类可能性(二项式检验)】选项,单击【选项】按钮,弹出如图 12-18 所示的【二项式选项】对话框。



图 12-17 【设置】选项卡



图 12-18 【二项式选项】对话框

在【假设比例】输入框中,输入“成功”期望的概率值( $p$ )。该值需大于 0 且小于 1。默认值为 0.5。

在【置信区间】栏中,可以选择用以下方法计算二分类数据的置信区间:



- **【Clopper-Pearson (精确)】**。它是基于累积二项式分布计算得到的精确区间。
- **【Jeffreys】**。它是基于  $p$  的后验分布且应用 Jeffreys 先验概率计算得到的 Bayesian 区间。
- **【似然比】**。它是基于  $p$  的似然函数计算得到的区间。

在**【定义分类字段的成功值】**栏中，可以指定如何为分类字段定义检验“成功”数据值的对照假设的比例。

- **【使用在数据中找到的第一个类别】**。它将使用在样本中找到的第一个定义“成功”的值来执行二项式检验。此选项仅适用于只有两个值的名义或有序字段。如果使用了此选项，则在**【字段】**选项卡中指定的所有其他分类字段都不会检验。这是默认值。
- **【指定成功值】**。使用指定值去定义“成功”的值列表来执行二项式检验。可以指定字符串或数值列表。列表中的值不一定要在样本中出现。

**【定义连续(型)字段的成功值】**栏。成功被定义为等于或小于割点的值。割点可以选择：

- **【样本中点】**。用最小值和最大值的平均值作为设置的分割点。
- **【定制割点】**。允许用指定的一个值作为分割点。

单击**【确定】**按钮，返回**【设置】**选项卡。

单击**【运行】**按钮，则在输出窗中得到运行结果。

② **【卡方检验】**。卡方检验可以应用到名义和有序变量。这将生成一个单样本检验，它可以根据变量类别的观察和期望频率间的差异来计算卡方统计量。

选择**【比较观察可能性和假设可能性(卡方检验)】**选项，单击**【选项】**按钮，弹出如图 12-19 所示的对话框。

在**【选择检验选项】**框中，有两个选项：

- **【所有类别概率相等】**。将在样本中的所有类别间生成均等的频率。这是系统默认值。
- **【自定义期望概率】**。允许为指定的类别列表指定不相等的频率。可以指定字符串或数值列表。列表中的值不需要在样本中出现。在类别列中指定类别值。在相对频率列中，为每个类别指定一个大于 0 的值。自定义的频率被视为比率，例如，指定频率 1、2 和 3 等同于指定频率 10、20 和 30，两者均指定了期望 1/6 的记录(即观测)属于第一个类别，1/3 的记录(即观测)属于第二个类别，1/2 的记录(即观测)属于第三个类别。在指定自定义期望概率时，自定义类别值必须包括数据中的所有变量(字段)值，否则将不对该变量(字段)执行检验。

单击**【确定】**按钮，返回**【设置】**选项卡。

单击**【运行】**按钮，则在输出窗中得到运行结果。

③ **Kolmogorov-Smirnov 检验**。Kolmogorov-Smirnov 检验可以应用在连续型变量(字段)上。它将生成一个单样本检验，即变量的样本累积分布函数是否为相同的均匀分布、正态分布、泊松分布或指数分布。

选择**【检验观察分布和假设分布(Kolmogorov-Smirnov 检验)】**选项，单击**【选项】**按钮，弹出如图 12-20 所示的选项卡。

在**【假设分布】**框下，有 4 种分布可供检验，它们分别是**【正态分布】**、**【均匀分布】**、**【指数分布】**和**【泊松分布】**。

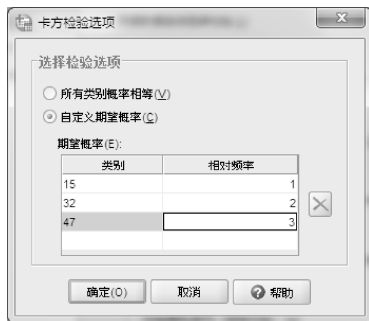


图 12-19 **【卡方检验选项】**对话框

- **【正态分布】**。在其**【分布参数】**栏中，可以**【使用样本数据】**，则使用样本中观察到的均值和标准差作为正态分布的参数。**【定制】**，则可以自定义参数值，在**【平均值】**框中输入所要检验的均值，在**【标准差】**框中输入所要检验的标准差值。
- **【均匀分布】**。在其**【分布参数】**栏中，可以**【使用样本数据】**，则使用样本中观察到的最小值和最大值作为均匀分布的参数。选择**【定制】**，则可以自定义参数值，在**【最小值】**框中输入所要检验的最小值，在**【最大值】**框中输入所要检验的最大值。
- **【指数分布】**。在其**【平均值】**栏中，可以**【使用样本数据】**，则使用样本中观察到的均值作为指数分布的参数。选择**【定制】**，则可以自定义参数值，在**【平均值】**框中输入所要检验的均值。
- **【泊松分布】**。在其**【平均值】**栏中，可以**【使用样本数据】**，则使用样本中观察到的均值作为泊松分布的参数。选择**【定制】**，则可以自定义参数值，在**【平均值】**框中输入所要检验的均值。

单击**【确定】**按钮，返回**【设置】**选项卡。

单击**【运行】**按钮，则在输出窗中得到运行结果。

④ **【Wilcoxon 符号秩检验】**。Wilcoxon 符号秩检验适用于连续型变量。

在**【设置】**选项卡中，选择**【比较中位数和假设中位数 (Wilcoxon 符号秩检验)】**选项，将生成一个关于变量中位数值的是否等于一个假设的中位数值的双样本检验。

作单样本的中位数检验时，应在该选项下面的**【假设中位数】**框中指定一个数值作为假设的中位数。

⑤ **【游程检验】**。虽然游程检验只适用于标记变量(只有两个类别的分类变量)，但可通过使用定义组别的规则而应用于所有变量。

选择**【检验随机序列 (游程检验)】**选项，将生成一个单样本检验，即二分变量的值序列是否为随机序列检验。单击**【选项】**按钮，弹出如图 12-21 所示的**【游程检验选项】**对话框。

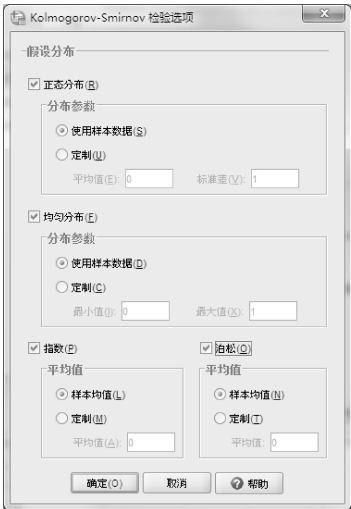


图 12-20 **【Kolmogorov-Smirnov 检验选项】**对话框

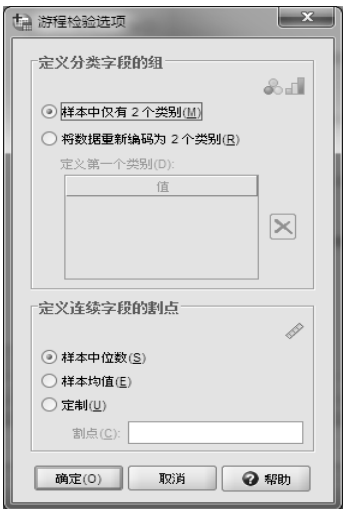


图 12-21 **【游程检验选项】**对话框

如果是分类数据，则在**【定义分类字段的组】**栏中进行选择。它在定义分类字段的组框中，有两个选项。如果选择**【样本中仅有 2 个类别】**选项，则使用在定义组的样本中找到的值来

执行游程检验。此选项仅适用于只有两个值的名义或有序变量。如果使用了此选项,则在【字段】选项卡中指定的所有其他分类变量都不会检验。如果选择【将数据重新编码为 2 个类别】选项,则使用自定义的某个组的值列表来执行游程检验。样本中的所有其他值定义其他组。列表中的值不需要在样本中出现,但每个组中必须至少有一条记录。

如果是连续型数据,则在【定义连续字段的割点】栏中进行选择。可以指定如何为连续型变量定义组。第一组定义为等于或小于割点的值。它可以通过以下 3 种方式来定义分割点:

- 【样本中位数】。在样本的中位数处设置分割点。
- 【样本均值】。在样本均值处设置分割点。
- 【定制】。自定义一个值作为分割点。在【割点】框中输入一个指定的值。

单击【确定】按钮,返回【设置】选项卡。

单击【运行】按钮,则在输出窗中得到运行结果。

### (2) 检验选项。

单击【设置】选项卡【选择项目】框中的【检验选项】,则得到如图 12-22 所示的【检验选项】选项卡,可以设定显著性水平和置信度,还可以选择如何处理含有缺失值的样品。

①【显著性水平】框。可以指定所有检验的显著性水平的  $\alpha$  值。它应介于 0~1 之间。系统默认值为 0.05。

②【置信区间(%)】框。可以指定所有生成置信区间的置信度。它应介于 0~100 之间。系统默认值为 95。

③【已排除的个案】栏。本栏中共有两个选项,可以用来确定参与检验的样品。



图 12-22 【检验选项】选项卡

- 【按检验排除个案】。在指定检验中,把此检验中所使用变量里含有缺失值的记录(观测)排除在检验之外。如果在分析中指定了多个检验,则将分别独立计算每个检验。
- 【按列表排除个案】。在所有分析中,把在【字段】选项卡选定的任何变量中含有缺失值的记录排除在检验之外。

### (3) 用户缺失值。

单击【设置】选项卡【选择项目】框中的【用户缺失值】选项,则得到如图 12-23 所示的【用户缺失值】选项卡。在【分类字段的用户缺失值】栏中可选取对缺失值的处理方式。要在分析中包含记录(观测),则对于分类变量来说,它必须具有有效值。

①【排除】。在分析中不包含用户缺失值。

②【包括】。在分析中包含用户缺失值。

不管怎样,对于系统缺失值和连续型变量中的缺失值,无论选取哪个选项,它都被视为无效。

## 12.9.1.3 单样本检验的实例分析

【例 10】 试检验【例 2】中的 100 名健康成年女子血清总蛋白含量是否服从均值为 7.35、

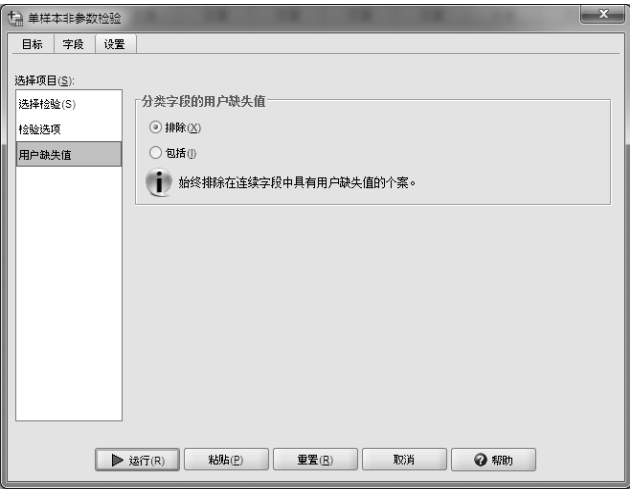


图 12-23 【用户缺失值】选项卡

不管怎样，要用样本数据检验其隶属的总体是否服从正态分布，或是否服从已知总体参数的正态分布，均可以在【非参数检验】子菜单的【单样本】过程的设置中，选择【检验观察分布和假设分布(Kolmogorov-Smirnov 检验)】选项来进行检验。不同的只是在两种情况下，一种是在假设正态分布的分布参数选项中选择【使用样本数据】选项，另一种是选择【定制】选项并输入相应的均值和标准差。

在 SPSS 中进行本例的分析步骤如下：

(1) 按【分析→非参数检验→单样本】顺序打开【单样本非参数检验】的【目标】选项卡，在【您的目标是什么？】栏中选择【自定义分析】选项。

(2) 单击【字段】按钮，打开【单样本非参数检验】的【字段】选项卡，选择【使用定制字段分配】选项，在【字段】列表框中选血清蛋白含量变量，单击右移按钮，将其移入【检验字段】框中。

(3) 单击【设置】按钮，打开【单样本非参数检验】的【设置】选项卡，选择【自定义检验】选项，选中【检验观察分布和假设分布(Kolmogorov-Smirnov 检验)】选项，单击该选项下的【选项】按钮，在【Kolmogorov-Smirnov 检验选项】对话框中选择【正态分布】选项，在【分布参数】选项中选择【定制】，在【平均值】框中输入“7.35”，在【标准差】框中输入“0.39”，单击【确定】按钮，返回【设置】选项卡。

(4) 单击【运行】按钮，则在输出窗中得到表 12-24 所示的假设检验的汇总表。

(5) 输出表说明。在原假设栏中，列出了完整的原假设；在测试(应为检验)栏中，列出了检验方法名称；在 Sig. 中，列出了  $P$  值( $P=0.711>0.05$ )；而在决策者栏中，给出了统计检验结果：保留原假设。

(6) 结论：现有证据表明，没有充分的理由可以拒绝原假设，故不拒绝原假设。

(7) Kolmogorov-Smirnov 检验中的其他信息。双击输出窗中的假设检验汇总表，则在弹出的模型浏览器中，得到如图 12-24 所示的详细结果。图中有 100 个数据资料按 0.5 组距所作的频数分布的直方图及其曲线图。此外，在直方图下面的表中还详细列出了 Kolmogorov-

标准差为 0.39 的正态分布。对应数据文件为 data12-10。

在【描述统计】子菜单的【探索分析】过程中，我们曾通过在【绘制】选项卡中选择【带检验的正态图】选项，使用其中的【Kolmogorov-Smirnov 检验】来对单样本数据资料进行正态性检验，初看起来，本例与其十分相似，但在使用该模块进行的统计分析中，它是假定样本均值等于总体的均值，样本的标准差等于总体的标准差，所以与本例不同，因为本例中要检验单样本数据是否服从一个已知总体均值和标准差的正态分布。

表 12-24 Kolmogorov-Smirnov 检验结果

假设检验汇总			
	原假设	测试	Sig. 决策者
1	血清蛋白含量的分布为正态分布，平均值为 7.35，标准差为 0.39。	单样本 Kolmogorov-Smirnov 检验	.711 保留原假设。

显示渐进显著性。显著性水平是 .05。

Smirnov 检验中的所用各种统计量值及双侧检验的概率  $P$  值。它对表 12-24 作了更详细的补充说明。

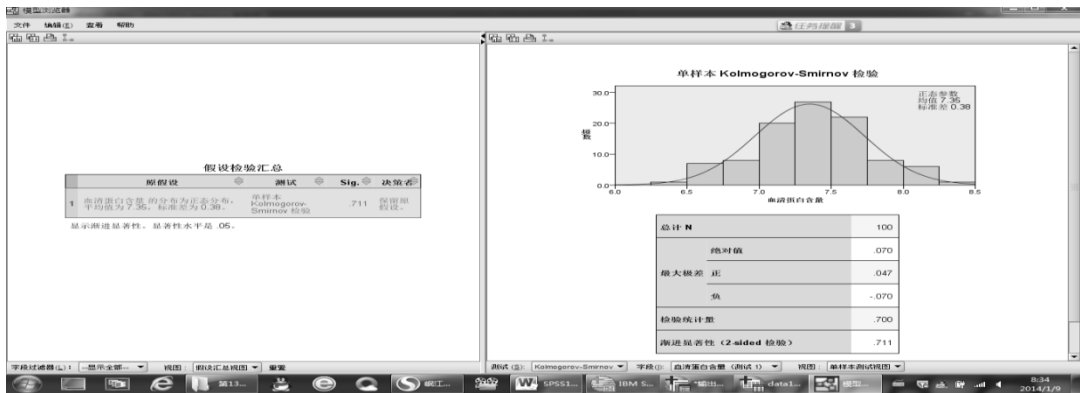


图 12-24 模型浏览器中显示的其他详细信息

## 12.9.2 独立样本检验

### 12.9.2.1 独立样本检验的用途

独立样本非参数检验使用一个或多个非参数检验来识别两个或更多独立样本间的差异。它不需要假定检验的数据资料呈正态分布。

### 12.9.2.2 独立样本检验的操作

按【分析→非参数检验→独立样本】顺序打开如图 12-25 所示的独立样本非参数检验的【目标】选项卡。

#### 1. 设定检验目标

在【您的目标是什么？】栏中，可快速指定常用的不同检验设置。它共有 3 个选项，选择其中之一，在描述栏中显示对该选项的描述与说明。

(1) 【自动比较不同组间的分布】。系统默认选项，对两个独立样本数据自动应用 Mann-Whitney U 检验，或对多个独立样本数据应用 Kruskal-Wallis 单因素 ANOVA 检验。

(2) 【比较不同组间的中位数】。使用中位数检验来比较在不同组间观察到的中位数。

(3) 【自定义分析】。当希望手动修改【设置】选项卡中的检验设置时，选择本选项。注意，如果随后在【设置】选项卡上更改了与当前选定目标不一致的选项，则会自动选择该选项。

#### 2. 设置检验变量和分组变量

单击【字段】按钮，显示如图 12-26 所示的独立样本非参数检验的字段选项卡，可指定所要检验的变量及分组变量。

(1) 【使用预定义角色】。使用现有的变量信息。【字段】选项卡支持可用于预先选择分析变量的预定义角色。

在 SPSS 中，变量所扮演的角色一般分为输入(如预测变量、自变量)、输出或目标(如因变量)、同时用作输入和输出、没有角色分配、分区及拆分等。

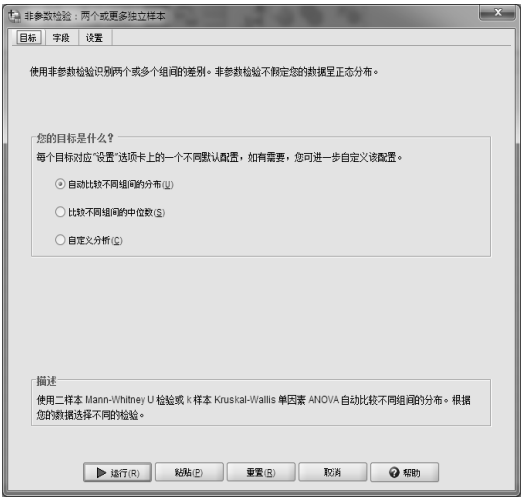


图 12-25 【非参数检验：两个或更多独立样本】对话框【目标】选项卡



图 12-26 【字段】选项卡

当打开【字段】选项卡时，满足角色要求的变量将自动显示在目标列表中，也就是【字段】框和【检验字段】框中。默认情况下，为所有变量分配输入角色。角色分配只影响支持角色分配的对话框，对命令语法没有影响。

所有预定义角色为“目标”或“两者”（变量将同时用作输入和输出）的变量将用作检验变量。如果有一个预定义角色为“输入”的分类变量，则它将用作分组变量；否则，默认不使用分组变量，必须【使用自定义字段分配】。至少需要一个检验变量和分组变量。

(2)【使用自定义字段分配】。设定的检验变量、分组变量可以代替其原变量角色。可以对【字段】框中列出的所要进行检验的变量移至【检验字段】框中进行指定，至少需选择一个变量。它可以避免【使用预定义角色】时，极可能将不需要检验的变量都选作检验变量的情况出现。

选择该选项后，可以对检验字段和分组字段作进一步指定：

- ①【检验字段】。在【字段】框中选择一个或多个连续型字段，将其移入【检验字段】框中。
  - ②【组】。在【字段】框中选择一个分类变量(用来分组的名义变量)，移入【组】框中。
- 在【字段】框中，用户可根据需要对出现在其中的变量列表按一定方式进行排序。在【排序】的下拉列表(见图 12-16)中，共有 3 种排序方式：

- ①【无】。系统默认方式。按它们在数据集中变量名出现的先后顺序进行排列。
- ②【字母数值】。按字母的 ASCII 码和数值大小顺序进行有序排列。
- ③【测量】(测度标准)。按测度标准即名义、有序、尺度顺序进行排序。在同一个测度标准中，则按它们在数据集中变量名出现的先后顺序进行排列。

单击【字段】列表框下面的全部(A)按钮，选中【字段】列表框中的所有变量；单击[ ]按钮，选中【字段】列表框中的所有名义变量；单击[ ]按钮，选中【字段】列表框中的所有有序变量；单击[ ]按钮，选中【字段】列表框中的所有尺度变量。

3. 设置检验方法及其选项

单击【设置】按钮，显示如图 12-27 所示的独立样本非参数检验的【设置】选项卡，可对在【字段】选项卡中所指定的变量需要执行的检验及其选项进行指定。



图 12-27 【设置】选项卡

在【选择项目】栏中，需对【选择检验】、【检验选项】和【用户缺失值】3 个方面分别进行设置。

(1) 选择检验。

在系统默认情况下，【设置】选项卡处于【选择检验】的状态。此时，可以设定对所指定【字段】执行何种检验。它有两个确定检验方法的选项：

【根据数据自动选择检验】。这是系统默认选项。对分组变量中只有两个组的数据资料使用 Mann-Whitney U 检验，而对有  $k$  个组的数据使用 Kruskal-Wallis 单因素 ANOVA 检验。

【自定义检验】。用户可根据各自的检验目的，自行在以下给定的检验项中设定所要执行的检验。

①【比较不同组间的分布】栏。如果要检验不同组间的样本是否来自相同的总体，则选择本栏中的检验项。

- 【Mann-Whitney U(二样本)】检验。使用两组合并后所得到的每个样品的秩来检验两组是否来自同一个总体。分组变量中按升序排列的第一个值定义第一个组，第二个值定义第二个组。如果分组变量有两个以上的值，则不能使用本检验。
- 【Kolmogorov-Smirnov(二样本)】检验。对两个分布之间的中位数、离散度、偏度等的任何差异很敏感。如果分组变量有两个以上的值，则不能使用此检验。
- 【检验随机序列(二样本 Wald-Wolfowitz)】检验。以合并两样本数据排序后的分组值所组成的新样本成员为依据，再使用游程检验来检验新样本的随机性。如果分组变量有两个以上的值，则不能使用此检验。
- 【Kruskal-Wallis 单因素 ANOVA(k 样本)】检验。它是 Mann-Whitney U 检验的扩展，用于单因素 K 水平的非参数的单因素方差分析。根据用户需要，可进行  $k$  样本的多重比较。

选择本选项后，还可通过单击【多重比较】下拉列表对如何进行多重比较作出选择。选择【无】，则不作多重比较；选择【所有成对比较】，则进行两两样本间的比较，这也是系统的默认选项；选择【逐步降低】，则以分组值最大的那个组为对照组，分别与其他各组进行两两比较。

- 【检验有序选项(k 样本 Jonckheere-Terpstra)】检验。其功能比 Kruskal-Wallis 检验更加

强大,但前提条件是  $k$  样本需具有自然顺序。因此,当  $k$  个总体有序(升序或降序)时,此检验方法非常有效。例如, $k$  个总体可以描述  $k$  个增加的温度。检验的假设是不同的温度产生同样反应的分布,备择假设:温度升高反应剧烈。这里,假设两个样本是有序的,因此,使用 Jonckheere-Terpstra 检验是最适当的。

选择本项后,还需要指定检验的顺序。共有两个选项:选择【最小到最大】,则规定其他假设:第一组的位置参数不等于第二组的,第二组的分量参数又不等于第三组的,依此类推;选择【最大到最小】,则规定其他假设:最后一组的位置参数不等于倒数第二组的位置参数,倒数第二组的又不等于倒数第三组的,依此类推。此外,还可通过单击【多重比较】下拉列表对如何进行多重的比较作出选择。其他含义同上。

②【比较不同组间的范围】栏。如果要检验不同组间的样本是否具有相同的范围,则选择本栏中的【Moses 极端反应】检验项。

【Moses 极端反应(二样本)】检验用来检验控制组与比较组是否具有相同范围。分组变量中按升序排列的第一个分组值定义控制组,第二个值定义比较组。如果分组字段有两个以上的值,则不能使用此检验。

选择本检验需要定义样本的离群值,可通过下面两个选项之一来完成:

- 【计算样本离群值】。计算样本两端 5% 的样品作为样本的离群值处理。
- 【离群值的定制数量】。在【离群体】框中输入一个正整数,表示将被作为离群值的数量。系统默认值为 1。

③【比较不同组间的中位数】栏。如果要检验不同组间的样本是否具有相同的中位数,则选择本栏中的【中位数检验( $k$  样本)】检验项。

选择本检验需要定义样本的中位数,可通过下面两个选项之一来完成:

- 【汇聚样本中位数】。用合并样本后计算得到的中位数作为它们共同的中位数。
- 【定制】。在【中值】框中输入一个值作为假设中位数。

此外,在本栏中还可根据需要进行多重比较分析。在系统默认情况下,将作所有成对比较;如果在【多重比较】下拉列表中选择【无】,则不作多重比较;如果选择【逐步降低】,则以分组值最大的那个组为对照组,分别与其他各组进行两两比较。

④【估计不同组间的置信区间】栏。如果要给出两个独立样本估计的置信区间,则选择本栏中的【Hodges-Lehman 估计(二样本)】选项。如果分组变量有两个以上的值,则不能使用此检验。单击【运行】按钮,则在输出窗中得到运行结果。

## (2) 检验选项。

单击【设置】选项卡的【选择项目】框中的【检验选项】,则得到如图 12-22 所示的【检验】选项卡。在【检验】选项卡上,可以设定显著性水平和置信度,还可以选择处理含有缺失值的样品的方法。

① 在【显著性水平】后框中,可以指定所有检验的显著性水平  $\alpha$  的值。它应介于 0 和 1 之间。系统默认值为 0.05。

②【置信区间(%)】。在其后框中,可以指定所有生成置信区间的置信度。它应介于 0 和 100 之间。系统默认值为 95。

③【已排除的个案】栏。在本栏中共有两个选项,可以用来确定参与检验的样品。

- 【按检验排除个案】。在指定的检验中,把此检验中所使用变量里含有缺失值的记录(观测)排除在检验之外。如果在分析中指定了多个检验,则将分别独立计算每个检验。



- **【按列表排除个案】**。在所有分析中，将把在字段选项卡选定的任何变量中含有缺失值的记录(观测)排除在检验之外。

(3) 用户缺失值。

单击**【设置】**选项卡的**【选择项目】**框中的**【用户缺失值】**选项，则得到如图 12-23 所示的**【用户缺失值】**选项卡。

在该选项卡的**【分类字段的用户缺失值】**栏中，选取对缺失值的处理方式。对分类变量而言在分析中所要包含的观测中需均为有效值。

- ① **【排除】**。在分析中不包含用户缺失值。
- ② **【包括】**。在分析中包含用户缺失值。

不管怎样，对于系统缺失值和连续型变量中的缺失值，无论选取哪个选项，它都被视为无效。在分析时被剔除。

12.9.2.3 独立样本检验的实例分析

**【例 11】**仍以**【例 7】**为例，对应数据文件为 data12-07，变量 ydp1 为优等品率，尺度测度变量；ff 为操作方法，名义变量，1 表示操作方法 1，2 表示操作方法 2，3 表示操作方法 3，4 表示操作方法 4。由于种种操作方法彼此独立，故用非参数检验中的独立样本检验过程来检验操作方法对产品的优等品率是否有显著影响。

操作步骤如下：

(1) 打开数据文件 data12-07，按**【分析→非参数检验→独立样本】**顺序打开独立样本非参数检验的**【目标】**选项卡。

(2) 在**【您的目标是什么？】**栏中，选择系统默认选项**【自动比较不同组间的分布】**。这意味着将对本例中具有多个独立样本的数据应用 Kruskal-Wallis 单因素 ANOVA 检验。

(3) 单击**【字段】**按钮，在**【字段】**选项卡中选择**【使用定制字段分配】**选项，将**【字段】**框中的 ydp1 变量移到**【检验字段】**框中，将**【字段】**框中的 ff 变量移到**【组】**框中。其他使用系统默认选项。

(4) 单击**【运行】**按钮，则在输出窗中得到表 12-25 所示的计算结果。

(5) 输出表说明。在原假设栏中，列出了完整的原假设，在**【测试(应为检验)】**栏中，列出了检验方法名称；在 Sig. 中，列出了  $P$  值 ( $P = 0.009 < 0.05$ )；而在决策者栏中，给出了统计检验结果：拒绝原假设。

(6) 结论：现有证据表明，有充分的理由可以拒绝原假设，说明操作方法对产品的优等品率有显著影响。

(7) Kruskal-Wallis 检验中的其他信息。双击输出窗中的假设检验汇总表，则在弹出的模型浏览器中，得到如图 12-28 所示的详细结果，上半部分为箱图，显示了 4 个组的观测值的分布情况，图中实体部分是正常值的分布范围，中间的黑体粗线为各组的中位数所在位置，实体部分的上方或下方出现的单值为异常值。箱图下面的表中还详细列出了 Kruskal-Wallis 检验所用到的各种统计量值及其双侧检验的概率  $P$  值，它对表 12-25 作了更详细的补充说明。

如图 12-29 所示，单击模型浏览器右下方**【测试】**(检验)下拉列表可见本次检验方法的名

表 12-25 Kruskal-Wallis 检验结果

假设检验汇总				
	原假设	测试	Sig.	决策者
1	ydp1 的分布在 ff 类别上相同。	独立样本 Kruskal- Wallis 检验	.009	拒绝原 假设。

显示渐进显著性。显著性水平是 .05。

称；单击【字段】下拉列表可见本次检验中所用到的变量名；单击【视图】下拉列表，可以进一步选择其中的选项来观察输出中的其他信息。

单击【视图】下拉列表中的【分类字段信息】选项，得到如图 12-30 所示的各分类的直方图。它显示了各类中的样本量及其在总观测值中所占的百分比。

单击【视图】下拉列表中的【连续字段信息】选项，得到对连续变量(检验变量)划分区间以后观测到的落入各区间的观测值数的直方图(见图 12-31)，它还包括总观测值量、最大值、最小值、均值、标准差等信息。

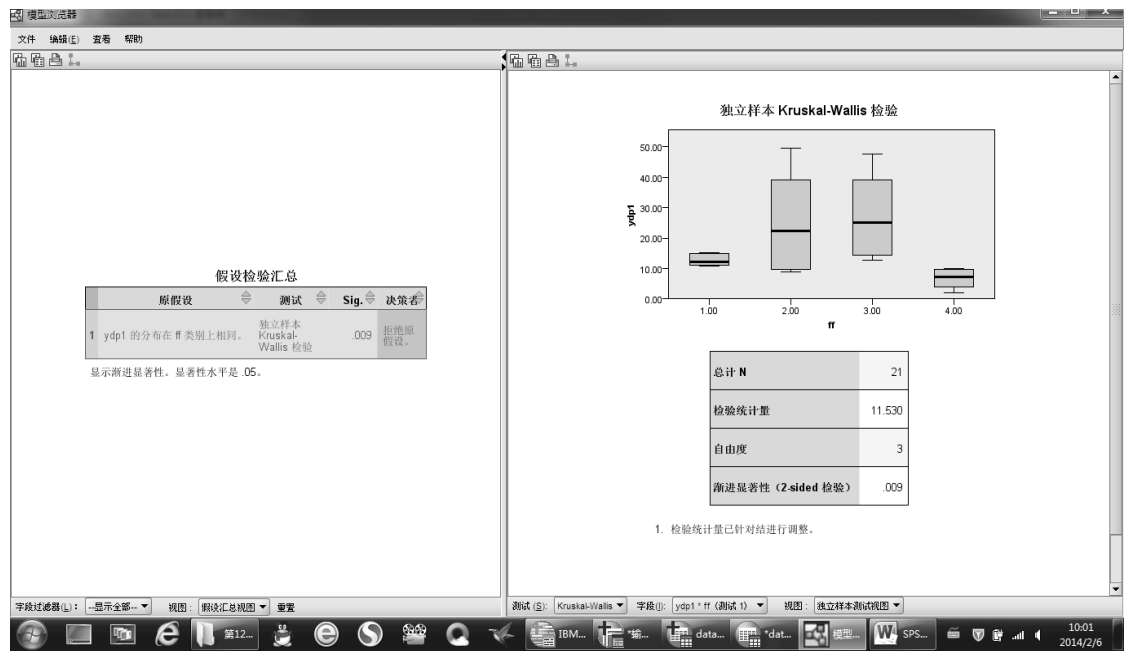


图 12-28 模型浏览器中显示的其他详细信息

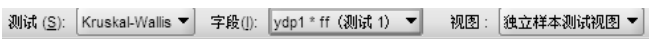


图 12-29 模型浏览器的下拉列表

单击【视图】下拉列表中的【成对比较】选项，得到如图 12-32 所示的各样本的平均秩的图示及如图 12-33 所示的多重比较结果表。从图 12-32 可见，第 1 组的平均秩为 10.40，第 2 组的平均秩为 13.75，第 3 组的平均秩为 15.80，第 4 组的平均秩为 3.50。从图 12-33 所示的多重比较结果表中可见，第 4 组与第 2 组(调整  $P = 0.038 < 0.05$ )、第 4 组与第 3 组(调整  $P = 0.010 < 0.05$ )的平均值之间的差异均有显著性意义。

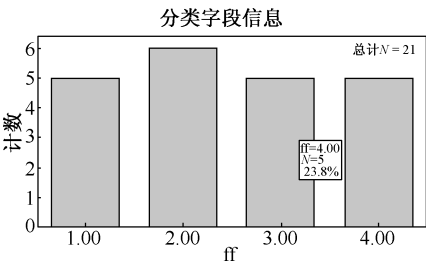


图 12-30 各分类的直方图

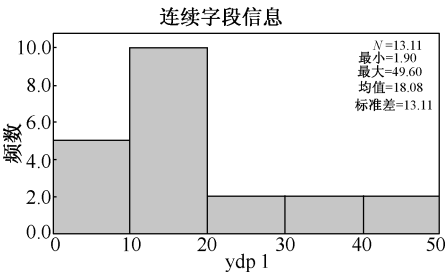
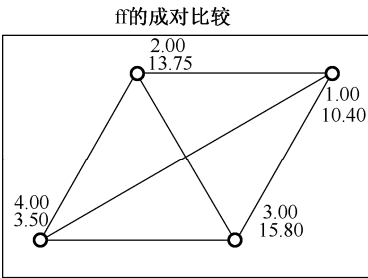


图 12-31 检验变量的直方图



每个节点显示n的样本平均秩

图 12-32 各样本的平均秩

样本1-样本2	检验统计量	标准误	标准检验统计量	Sig.	调整显著性
4.00-1.00	6.900	3.923	1.759	.079	.472
4.00-2.00	10.250	3.756	2.729	.006	.038
4.00-3.00	12.300	3.923	3.135	.002	.010
1.00-2.00	-3.350	3.756	-.892	.372	1.000
1.00-3.00	-5.400	3.923	-1.376	.169	1.000
2.00-3.00	-2.050	3.756	-.546	.585	1.000

每行检验原假设：样本 1 和样本 2 分布相同。  
显示渐进显著性（2-sided 检验）。显著性水平是 .05。

图 12-33 多重比较

12.9.3 相关样本检验

12.9.3.1 相关样本检验的用途

相关样本非参数检验使用一个或多个非参数检验来识别两个或更多相关样本间的差异。它不需要假定检验的数据资料呈正态分布。

相关样本与独立样本有明显的区别：在独立样本中，同一个样本里的任意一个变量中的观测值的先后顺序可以随意改变，而不影响最终的分析；但在相关样本中，存储在数据集变量中的两个及两个以上的相关的测量值均须在同一个被试对象的记录中，这也就是说是不可以单独改变其中任一个变量中的观测值的先后顺序的。

例如，如果用定期间隔测试的方式获取每个被试者的体重并存储在“节食前体重”、“中间体重”和“节食后体重”这样的变量中，则可使用相关样本的非参数检验分析来研究节食计划的有效性。这些变量称为相关变量。

12.9.3.2 相关样本检验的操作

按【分析→非参数检验→相关样本】顺序打开如图 12-34 所示的相关样本非参数检验的【目标】选项卡。

1. 设定检验目标

在【您的目标是什么?】栏中，可快速指定常用的不同检验设置。它共有两个选项：

(1) 【自动比较观察数据和假设数据】。这是系统默认选项。当只指定两个变量，且为分类数据时，软件自动使用 McNemar 检验；当指定两个以上变量，且为分类数据时，软件自动使用 Cochran 的 Q 检验；当只指定两个变量，且为连续数据时，软件自动使用 Wilcoxon 匹配对符号秩检验；当指定两个以上变量，且为连续数据时，软件自动使用 Friedman 双因素 ANOVA 非参数检验。

(2) 【自定义分析】。当希望手动修改【设置】选项卡中的检验设置时，选择本选项。如果随后在【设置】选项卡上更改了与当前选定目标不一致的选项，则会自动选择该设置。

注意：当指定了不同测度水平的变量时，软件将首先用测量水平对变量进行区分，然后对各个组将使用相应的检验。例如，如果用户选择【自动比较观测数据和假设数据】作为目标，并指定 3 个连续型变量和 2 个名义变量，那么软件将自动会对连续变量使用 Friedman 检验，并对名义变量使用 McNemar 检验。



图 12-34 【非参数检验：两个或更多相关样本】对话框【目标】选项卡

2. 设置检验字段和分组变量

单击【字段】按钮，显示如图 12-35 所示的相关样本非参数检验【字段】选项卡。



图 12-35 【字段】选项卡

在【字段】选项卡中，指定要对哪些变量进行检验。

(1) 【使用预定义角色】。此选项使用现有的变量信息。【字段】选项卡支持可用于预先选择分析变量的预定义角色。

在 SPSS 中，变量所扮演的角色一般分为输入(如预测变量、自变量)、输出或目标(如因变量)、同时用作输入和输出、没有角色分配、分区及拆分等。

当打开【字段】选项卡时，满足角色要求的变量将自动显示在目标列表中。默认情况下，为所有变量分配输入角色。角色分配只影响支持角色分配的对话框。




所有预定义角色为“目标”或“两者”(变量将同时用作输入和输出)的变量将用作检验字段。至少需要两个检验变量。

(2) 【使用自定义字段分配】。设定的检验变量可以代替其原变量角色。

选定该选项后,可以对检验变量作进一步指定。在【字段】框中选择两个或多个连续型变量,将其移入【检验字段】框中。每个变量对应一个单独的相关样本。

在【字段】框中,根据需要,用户可对出现在其中的字段列表按一定方式进行排序。在字段排序的下拉列表(见图 12-16)中,共有 3 种排序方式:

- ① 【无】。系统默认方式。按它们在数据集中变量名出现的先后顺序进行排列。
- ② 【字母数值】。按字母的 ASCII 码和数值大小顺序进行有序排列。
- ③ 【测量】(测度标准)。按测度标准即名义、有序、尺度顺序进行排序。在同一个测度标准中,则按它们在数据集中变量名出现的先后顺序进行排列。

单击字段名列表框下面的 **全部(A)** 按钮,则选中字段名列表框中的所有字段。单击  按钮,则选中字段名列表框中的所有名义字段。单击  按钮,则选中字段名列表框中的所有有序字段。单击  按钮,则选中字段名列表框中的所有尺度字段。

### 3. 设置检验方法及其选项

单击【设置】按钮,弹出如图 12-36 所示的相关样本非参数检验【设置】选项卡,可对在【字段】选项卡中指定的【字段】需要执行的检验及其选项进行指定。

在【选择项目】栏中,需对选择检验、检验选项和用户缺失值 3 个方面进行设置。

(1) 选择检验。

在系统默认情况下,【设置】选项卡界面处于【选择检验】的状态。此时,可以设定对所指定变量执行何种检验。它有两个确定检验方法的选项:

【根据数据自动选择检验】选项。这是系统默认选项。当只指定两个变量时,软件自动使用 McNemar 检验;当指定两个以上分类变量时,软件自动使用 Cochran 的 Q 检验;当只指定两个连续型变量时,软件自动使用 Wilcoxon 匹配对符号秩检验;当指定两个以上连续型变量时,软件自动使用 Friedman 双因素 ANOVA 非参数检验。

【自定义检验】选项。用户可根据各自的检验目的,自行在以下给定的检验项中设定所要执行的检验。

① 【检验二分类数据中的更改(变化)】栏。

- 【McNemar 检验(二样本)】。McNemar 检验可用来确定初始的响应率(事件前)是否等于最终响应率(事件后)。如果是二分数据,每个被试对象的响应分别在指定事件发生的前、后被重复测定,则可选择 McNemar 检验。本检验对于在前后对比设计中检测由实验干预引起的响应变化很有用。
- 【Cochran Q(k 样本)】检验。如果所有的响应均是二值的情形,那么在要检验  $k$  个相关

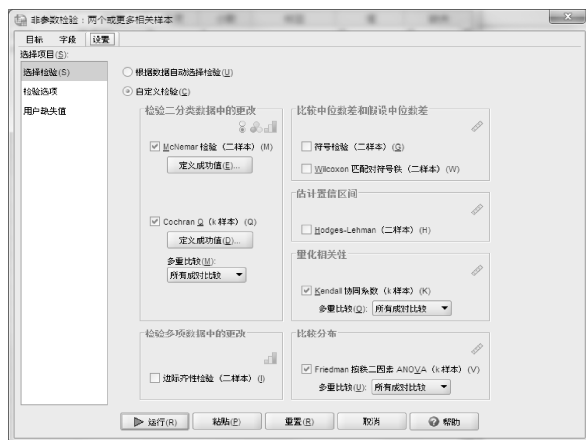


图 12-36 【设置】选项卡

样本有相同的均数时,请选择本选项。选中本项后,【定义成功值】按钮和【多重比较】下拉菜单将被激活。

单击【定义成功值】,弹出如图 12-37 所示的【CochranQ: 定义成功值】对话框。

在【定义分类字段的成功】栏中可以指定如何为分类字段定义“成功”。



图 12-37 【Cochran Q: 定义成功值】对话框

- **【在数据中找到的第一个值】**。将使用在样本中找到的第一个值来定义“成功”,以此来执行检验。本选项仅适用于只有两个值的名义或有序字段;如果使用了本选项,则在【字段】选项卡中指定的所有其他分类字段都不会被检验。它是系统默认选项。
- **【将值合并为成功类别】**。将使用用户指定的“成功”值列表来执行检验。可以用字符串或数值列表来指定“成功”值。列表中的值不需要在样本中出现。

单击【确定】按钮,返回【设置】选项卡。

选择本选项后,还可通过单击【多重比较】下拉列表对如何进行多重比较作出选择。选择【无】,则不作多重比较;选择【所有成对比较】,则进行两两样本间的比较,这是系统的默认选项;选择【逐步降低】,则以分组值最大的那个组为对照组,分别与其他各组进行两两比较。

② **【检验多项数据中的更改(变化)】**栏。本栏中只有一个**【边际齐性检验(二样本)】**选项。边际齐性检验通常在重复测量的情况下使用。它是 McNemar 检验从二值响应到多项响应的扩展,是一种用来检验配对有序变量的对应分类值之间出现的可能性是否相同的非参数检验方法。如果在【字段】选项卡上指定两个以中的变量将不执行本检验。

③ **【比较中位数差和假设中位数差】**栏。在本栏中有两个选项。用来检验两个连续字段间的中位数差值是否等于 0。如果在“字段”选项卡上指定两个以上的字段,将不执行这些检验。

- **【符号检验(二样本)】**。适用于两个相关样本的配对数据资料。
- **【Wilcoxon 配对符合秩(二样本)检验】**。适用于两个相关样本的配对数据资料。

④ **【估计置信区间】**栏。如果要给出两个相关样本的配对连续型变量之间的中位数差值估计的置信区间,则选择本栏中的**【Hodges-Lehman 估计(二样本)】**选项。如果在【字段】选项卡中指定两个以上的变量,将不执行此检验。

⑤ **【量化相关性(关联)】**栏。本栏中只有一个**【Kendall 协同系数(k 样本)】**选项。如果在裁判员或评判员对多个被试对象同时给出评分后,需要对裁判员或评判员之间的评分进行一致性评定时,可选择本选项。选定本选项后,【多重比较】下拉列表被激活,用户还可对如何进行多重比较作出选择。

单击【多重比较】下拉列表,选择【无】,则不作多重比较;选择【所有成对比较】,则进行两两样本间的比较,这也是系统的默认选项;选择【逐步降低】,则以分组值最大的那个组为对照组,分别与其他各组进行两两比较。

⑥ **【比较分布】**栏。本栏中只有一个**【Friedman 按秩二因素 ANOVA(k 样本)】**选项。如果要作  $k$  个相关样本是否来自同一总体的检验,则选择此项。

同样,在选定本选项后,【多重比较】下拉列表被激活,用户还可对如何进行多重比较作出选择。

单击【多重比较】下拉列表,选择【无】,则不作多重比较;选择【所有成对比较】,则进行两两样本间的比较,这也是系统的默认选项;选择【逐步降低】,则以分组值最大的那个

组为对照组，分别与其他各组进行两两比较。

单击【运行】按钮，则在输出窗中得到运行结果。

(2) 检验选项。

单击【设置】选项卡的【选择项目】框中的【检验选项】，则得到如图 12-22 所示的【检验】选项卡。在检验选项卡上，可以设定显著性水平和置信度，还可以选择如何处理含有缺失值的样品。

① 在【显著性水平】后框中，可以指定所有检验的显著性水平 $\alpha$ 的值。它应介于 0 和 1 之间。系统默认值为 0.05。

②【置信区间(%)】。在其后框中，可以指定所有生成置信区间的置信度。它应介于 0 和 100 之间。系统默认值为 95。

③【已排除的个案】栏。在本栏中共有两个选项，可以用来确定参与检验的样品。

- 【按检验排除个案】选项。在指定检验中，将把此检验中所使用变量里含有缺失值的记录(样品)排除在检验之外。如果在分析中指定了多个检验，则对每个检验分别处理。
- 【按列表排除个案】选项。在所有分析中，将把在字段选项卡选定的任何变量中含有缺失值的记录(样品)排除在检验之外。

(3) 用户缺失值

单击【设置】选项卡【选择项目】框中的【用户缺失值】选项，则得到如图 12-23 所示的【用户缺失值选项卡】对话框。

在该选项卡的【分类字段的用户缺失值】栏中，选取对缺失值的处理方式。在分析中包含的分类变量的记录，须为有效值。

- ①【排除】选项。选择此项，则在分析中不包含用户缺失值。
- ②【包括】选项。选择此项，则在分析中包含用户缺失值。

不管怎样，对于系统缺失值和连续型变量中的缺失值，无论选取哪个选项，它都被视为无效。

12.9.2.3 相关样本检验的实例分析

【例 12】 某村在村长选举前半年，用 1 表示认可，0 表示不认可，对 3 位村长候选人的认可度在随机抽取的 50 位村民中进行了摸底调查。调查结果存放在数据文件 data12-11 中。试问抽样调查结果是否可认为村民对 3 位候选人的认可度是一样的？

本例将使用非参数检验中的相关样本过程自动进行分析，因此，在运行之前对即将运行的数据文件作必要的说明。

数据文件中 3 个字段(变量)的属性见图 12-38。

注意：在测度水平(度量标准)中，一定要将变量的测度水平定义为名义测度；另外，在角色中要将变量定义为目标变量。这是让 SPSS 软件进行自动识别的关键。

	名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量标准	角色
1	候选人 1	数值(N)	8	0		{1. 认可}...	无	8	靠右	名义(N)	目标
2	候选人 2	数值(N)	8	0		{1. 认可}...	无	8	靠右	名义(N)	目标
3	候选人 3	数值(N)	8	0		{1. 认可}...	无	8	靠右	名义(N)	目标

图 12-38 数据文件中变量的属性

这些变量都是相关的二分名义变量，要进行村民对 3 位候选人的认可度的一致性分析，可用 Cochran 的 Q(k 样本)检验。

此外，需要记住的是，在数据文件中的第一个值(第一个名义变量的第一个观测值)为 1，而不是 0。

现在，要在 SPSS 中使用非参数检验中的相关样本过程进行 Cochran 的 Q(k 样本)检验的操作步骤非常简单，均采用系统默认选项即可。因此，只需要进行如下简单操作：

按【分析→非参数检验→相关样本】顺序打开相关样本非参数检验的【目标】选项卡，单击【确定】按钮运行，则在输出窗中得到表 12-26 所示的输出结果。

表 12-26 检验汇总  
假设检验汇总

	原假设	测试	Sig.	决策者
1	候选人1, 候选人2 and 候选人3 的分布相同。	相关样本 Cochran Q 检验	.005	拒绝原假设。

显示渐进显著性。显著性水平是 .05。

检验：在 Sig.中，列出了  $P$  值( $P=0.005<0.05$ )；而在决策者栏中，给出了统计检验结果：拒绝原假设，即村民对他们有不同的认可度。

那么，哪位候选人有较高的认可度呢？要知道这样的信息，需在输出窗中双击上表，在弹出的模型浏览器中，可得到有关这方面的详细信息，见图 12-39。

在模型浏览器右上方的直方图中显示了各位候选人所获得认可和不认可响应的人数及构成比情况。在默认情况下，使用在数据中找到的第一个类别来定义成功值，也就是使用在样本中找到的第一个值(已知样本中第一个值为 1)来定义“成功”，软件中“成功”值用 0 表示。因此，在该直方图中，图示 0 表示 1 所代表的类别，即认可；而图示 1 则代表另一个类别，即不认可。因此，有 82%的被调查的村民认可第 1 位候选人。这个比例值可以通过将鼠标移向模型浏览器中候选人所在的直方图来获取。

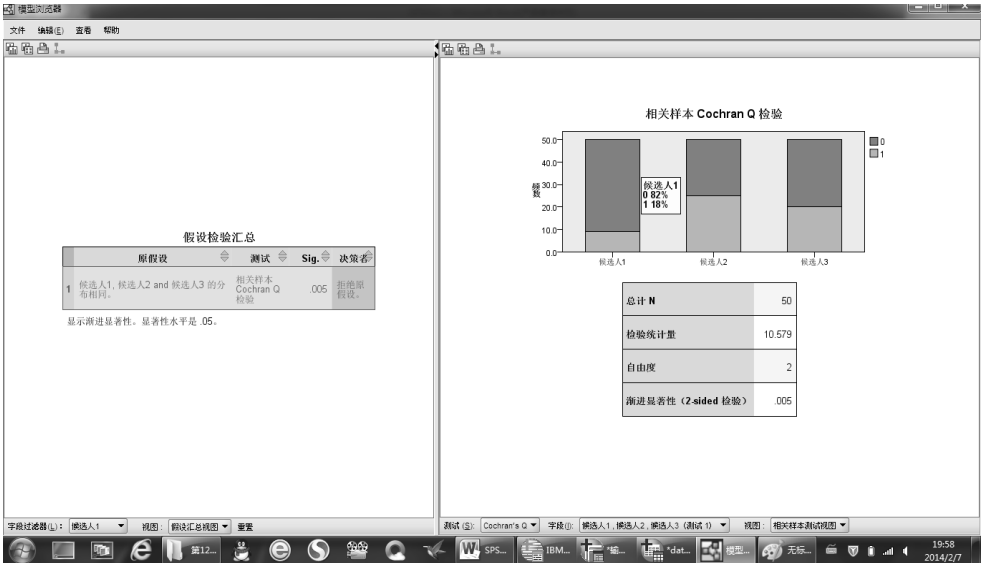


图 12-39 显示在模型浏览器中的图形

在模型浏览器右下方的检验表中显示了 Cochran Q 检验中相关统计量的计算结果。另外，在模型浏览器的【视图】下拉列表中，单击分类字段信息，再在【字段】下拉列表中选择相应的候选人选项，则可清晰地看到各位候选人得到认可和不认可分布情况的直方图，见图 12-40。



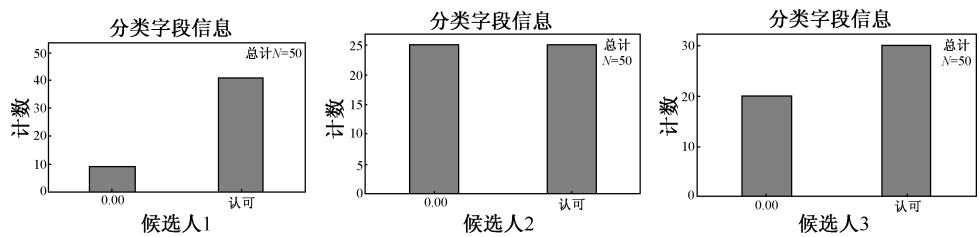


图 12-40 各位候选人得到村民的认可和不可分布情况的直方图

单击模型浏览器【视图】下拉列表中的【成对比较】，则得到如图 12-41 所示的成对比较图及表 12-27 所示的多重比较表。

由图 12-41 可见，第 1 位候选人得到的认可数为 41，第 2 位候选人得到的认可数为 25，第 3 位候选人得到的认可数为 30。

从表 12-27 可见，第 1 位候选人与第 2 位候选人的认可性之间有显著性差异 ( $p = 0.004$ )。

表 12-27 多重比较

样本1-样本2	检验统计量	标准误差	标准检验统计量	Sig.	调整显著性
候选人2-候选人3	-.100	.101	-.993	.321	.962
候选人2-候选人1	.320	.101	3.179	.001	.004
候选人3-候选人1	.220	.101	2.185	.029	.087

每行检验原假设：样本 1 和样本 2 分布相同。  
显示渐进显著性 (2-sided 检验)。显著性水平 = .05。



图 12-41 成对比较图

## 习 题 12

1. 什么是非参数检验？SPSS 的哪个过程可进行非参数检验？共包括几种方法？
2. 100 名健康成年女子血清蛋白含量记录在数据文件 data12-02 中，试用 Chi-Square 过程检验健康成年女子血清蛋白含量是否服从正态分布。
3. 对一台设备进行寿命试验，记录 10 次无故障工作时间，并从小到大排列在数据文件 data12-12 中。问此设备的无故障工作时间是否服从指数分布？
4. 一个监听装置收到的信号，记录在数据文件 data12-13 中，能否说该信号是纯粹随机干扰？
5. 两个地点的地表土壤 Ph 值记录在数据文件 data12-14 中，问这两个地点的平均 Ph 值是否一样。
6. 10 个病人进行某种药物疗法前、后的血压(收缩压，单位：毫米汞柱)记录在数据文件 data12-15 中，问该药物疗法是否有效。
7. 数据文件 data12-16 中是某村 20 个村民对 4 个候选人(A、B、C、D)的赞同与否的调查(数字 1 表示赞同，0 表示不赞同)数据，试用 Cochran's Q 法检验村民是否对这 4 个候选人有不同的看法。

# 第 13 章 聚类分析与判别分析

## 13.1 聚类分析、判别分析及其分析过程

分类学是人类认识世界的基础科学。聚类分析和判别分析是研究事物分类的基本方法，广泛地应用于自然科学、社会科学、工农业生产各个领域。

### 13.1.1 聚类分析

聚类分析(Cluster Analysis)是根据事物本身的特性研究个体分类的方法。聚类分析的原则是同一类中的个体有较大的相似性，不同类中的个体差异很大。

根据分类对象的不同，分为样品聚类和变量(又称指标)聚类。

#### 1. 样品聚类

样品聚类在统计学中又称为 Q 型聚类。用 SPSS 的术语来说就是对事件(或称样品或称观测)进行聚类，是根据被观测对象的各种特征，即反映被观测对象的特征的特征的各变量值进行分类的。例如，用 K-均值聚类分析，可以根据观众对电视机外观偏好的特点把电视机外观分为  $k$  组，并把该结果用于确定营销市场的分类，或把被调查的城市进行分类，以便对不同城市的策略进行比较。

应该注意的是，不同的目的选用不同的指标作为分类的依据。例如，为选拔少年运动员所选用的指标，就不同于课外活动小组所选用的指标；对啤酒按价格进行分类和按成分进行分类所选用的指标也是不同的。

#### 2. 变量聚类

变量聚类在统计学中又称为 R 型聚类。反映同一事物特点的变量很多，一般是根据所研究的问题选择部分变量对事物的某一方面进行研究。由于人类对客观事物的认识是有限的，往往难以找出彼此独立的有代表性的变量，从而影响了问题的进一步认识和研究。例如，在回归分析中，由于自变量的共线性导致偏回归系数不能真正反映自变量对因变量的影响。因此往往先要进行变量聚类，找出彼此独立且有代表性的自变量，而又不丢失大部分信息。在生产活动中也不乏需要进行变量聚类的实例。制衣业制定衣服型号就是根据人体各部分尺寸数据找出最有代表性的指标，如身長、胸围、裤长、腰围作为上衣和裤子的代表性指标；制鞋业中制定的鞋的型号也是如此。变量聚类使批量生产成为可能。

无论哪种聚类分析得出的结论都是为了某种目的所做的工作，有时并非在自然界真实存在这样的类。

### 13.1.2 判别分析

判别分析是根据表明事物特点的变量值和它们所属的类，求出判别函数，根据判别函数对未知所属类别的事物进行分类的一种分析方法。

在自然科学和社会科学的各个领域经常遇到需要对某个个体属于哪一类进行判断。例如，动物学家对动物分类的研究往往需要获得某个动物属于哪一科、目、纲等的判断，就可以根据判别分析已经得出的判别函数进行判断。

判别分析必须已知样品的所属类别。如果类别未知，必须先作聚类。根据样本聚类的结果进行判别分析，得出判别函数，进而对其他研究对象属于哪一类做出判断。例如，在选拔少年运动员时，首先要根据已有的少年运动员的身体形态、身体素质、心理素质、生理功能的各种指标(变量)进行测试，得到各种指标的测试值(变量值)，据此对少年运动员进行分类。根据分类结果再求出选材的判别函数，作为选材的依据。又如，可以根据啤酒中含有的酒精成分、钠成分及所含热量“卡路里”数值对啤酒进行分类。

判别分析与聚类分析不同点在于，判别分析要求已知一系列反映事物特征的数值变量的值，并且已知各个体的分类。

SPSS 中进行聚类分析和判别分析的统计分析过程，是由分析菜单中的分类命令导出的。如图 13-1 所示。二级菜单中是进行聚类分析、判别分析的过程清单。本章内容包括：

- (1) 两步聚类。是一个探索性的分析工具，可以分析大数据文件并自动确定最好的分析结果。
- (2) K-均值聚类。是一种快速聚类分析过程，仅对观测进行快速聚类。
- (3) 系统聚类。是分层聚类，进行样本聚类和变量聚类的过程。
- (4) 判别。进行判别分析的过程。

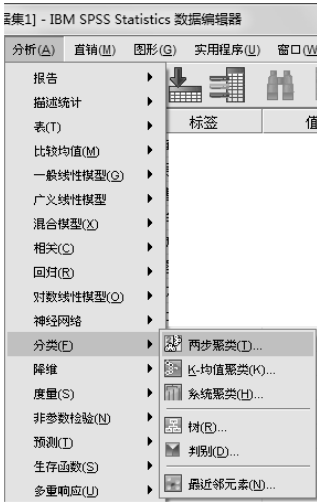


图 13-1 各种聚类分析过程

## 13.2 两 步 聚 类

### 13.2.1 两步聚类概述

#### 1. 两步聚类的概念

两步聚类过程是一个探索性的分析工具，为揭示自然的分类或分组而设计。这个过程所使用的算法有几个特色区别于传统的聚类分析技术。其特点是：分类变量和连续变量都可以参与两步聚类分析；可以自动确定分类数；可以高效率地分析大数据集；用户可以自己设置用于运算的内存容量。

两步聚类法在聚类过程中除了使用传统的欧式距离外，为了处理分类变量和连续变量，还使用似然距离测度，它要求模型中的变量是独立的。分类变量是多项式分布，连续变量是正态分布。虽然经验表明，参与分析的变量违反这一假设的情况下有时也可以得出结果，但还是应该使用其他 SPSS 过程检验参与分析的变量是否符合分类变量和连续变量在分布方面的要求。

可以使用双变量的相关过程去检验两个连续变量之间的独立性。使用交叉表过程检验两个分类变量之间的独立性；使用均值过程检验连续变量和分类变量之间的独立性；使用探索过程检验连续变量的正态性；使用卡方检验过程检验分类变量是否是多项式分布的。

所谓两步聚类就是，第一步对每个观测考察一遍，确定类中心。根据相近者为同一类的原则，计算距离并把与类中心距离最小的观测分到相应的各类中去。这个过程称作构建一个分类的特征树(CF)。首先，它把一个观测放在树的叶节点根部，该节点含有该观测的变量信息。然后，使用距离测度作为相似性的判据，每个后续的观测根据它与已经存在的节点的相似性归到某类中去。如果相似则将该观测加在一个已经存在的节点上，形成该节点的树叶；而如果不相似，就形成一个新节点。

第二步，使用凝聚算法对特征树的叶节点分组。凝聚算法可用来产生一个结果范围。为确定最好的类数，对每一个聚类结果使用 BIC 判据或 AIC 判据作为聚类判据进行比较，得出最后的聚类结果。

两步聚类过程的输出提供聚类得出结果的类数判据(AIC、BIC)、聚类最终结果的类频数等各类变量的描述统计量，可以产生类频数条形图、类频数饼图和变量重要性图。

2. 有关术语

(1) 聚类特征树。在聚类的第一步，根据计算的距离确定类结构。每类有一个节点，属于该类的观测就是该节点的树叶。由于树叶的不断增加构成树枝。第一步聚类过程就是 CF 树成长的过程。

(2) AIC 或 BIC 是在聚类的第二步凝聚过程中用到的两个判据，是两个算法，即 Akaik(AIC)判据或贝叶斯判据(BIC)。

(3) 调谐算法。两步聚类过程可以自动进行聚类，也可以人为控制聚类过程。在人为控制情况下，自己指定参数，称作调谐(Tuning)。参数指定了，CF 树的规模就基本确定了。

(4) 噪声处理。由于两步聚类要处理大数据集，在构建 CF 树时，如果指定了类数和算法的参数，如一个 CF 树最多的分枝数、一个叶节点最大子节点数等，那么在第一步聚类过程中，当观测很多时，就可能 CF 树满了，不能再长了。没有在树上的观测就称为噪声。对这些待处理的观测，用户可以调整算法参数，让 CF 树能容纳更多的观测，将其保留在某类中或者丢掉。这种处理称作噪声处理。

(5) 离群值。根据噪声处理参数聚类结束时被丢掉的观测称作离群值，单独构成一类，不计在聚类结果的类数中。

(6) 聚类质量的评判。根据 Kaufman 和 Rousseeuw(1990)关于聚类结构解释的研究成果用聚类结合和分离的 Silhouette 参数来判定结果的质量。良好的结果表示数据将 Kaufman 和 Rousseeuw 的评级反映为聚类结构的合理迹象或强迹象，尚可的结果将其评级反映为弱迹象，而较差的结果将其评级反映为无明显迹象。在质量图中使用不同颜色区分三等：较差、尚可或好。该图可以快速检查质量是否较差，如果较差，就可以返回建模节点修改聚类模型设置以生成较好的结果。

第  $i$  个观测的 Silhouette 的计算公式为

$$\text{Silhouette}_i = (B_i - A_i) / \max(A_i, B_i)$$

式中， $A_i$  是第  $i$  个观测到其聚类中心的距离； $B_i$  是观测到非所属但是最近聚类中心的距离。某类的参数 Silhouette 值就是该类各观测的该值的平均值。该值为 1 表示所有观测直接位于其类中心上；-1 表示所有观测位于某些其他类的类中心上；值为 0 表示在正常情况下观测到其本类中心与到最近其他类的类中心是等距的。

13.2.2 两步聚类过程

在使用两步聚类分析过程之前，应该对各变量的变量测度类型在 Variable Viewer 窗口中进行认

真的定义。并对各变量的独立性和分布特征进行检验。

两步聚类的操作步骤如下。

### 1. 打开主对话框

建立或读入数据文件后，按【分析→分类→两步聚类】顺序单击菜单项，打开如图 13-2 所示的对话框。

### 2. 在主对话框中

(1) 指定分析变量。左侧的源变量栏中显示了可以参加两步聚类分析的两种类型的变量。

① 选择参与聚类分析的分类变量，单击上面的向右箭头按钮，将其移到右侧的【分类变量】框中。两步聚类要对这个变量的值进行分类。

② 选择连续型变量，单击下面的向右箭头按钮，将其送入【连续变量】框中。两步聚类根据这些变量的值进行聚类。

(2) 【距离度量】栏。选择计算两类间的相似程度的算法。

① 【对数相似值】。该算法要求所有变量彼此独立，连续变量是正态分布的，分类变量是多项式分布的。

② 【Euclidean】。欧式距离法测度两类之间的“直线”距离。当所有参与聚类的变量都是连续变量时此方法才适用。

(3) 【聚类数量】栏。指定最后分类结果所分的类数。在该栏中指定要聚成几类。

① 【自动确定】类数。两步聚类过程用在聚类准则(判据)组中指定的判据，自动确定最好的类数。在【最大值】框中输入一个正整数，指定该过程应该考虑的最大类数。默认的最大类数是 15。最后的聚类结果，类数在 1 至指定的最大类数之间。

② 【指定固定值】。在【数量】框中输入一个正整数作为要求聚成的固定的类数。最后聚类结果必定是指定的类数。

(4) 【连续变量计数】栏。显示连续变量的计数。即在【二阶聚类：选项】对话框中指定要进行标准化的连续变量个数和假设已经标准化的连续变量的个数。

(5) 【聚类准则】栏。指定确定类数的判据，包括【施瓦茨的贝叶斯准则(BIC)】和【Akaike 信息准则(AIC)】。

### 3. 选项

单击【选项】按钮，打开如图 13-3 所示对话框。

(1) 【离群值处理】栏。选择在特征树满时，对还没有聚到任何一类中的离群观测值继续加入特征树的处理方法。如果类特征树满了，该组选项允许在聚类时对待分类的观测作特殊处理使 CF 树是完整的，如果不能接受更多的观测，在叶节点和非叶节点处可以分开。

【使用噪声处理】给出一个百分比。如果某节点包含的观测数与最大叶子数之比小于指定

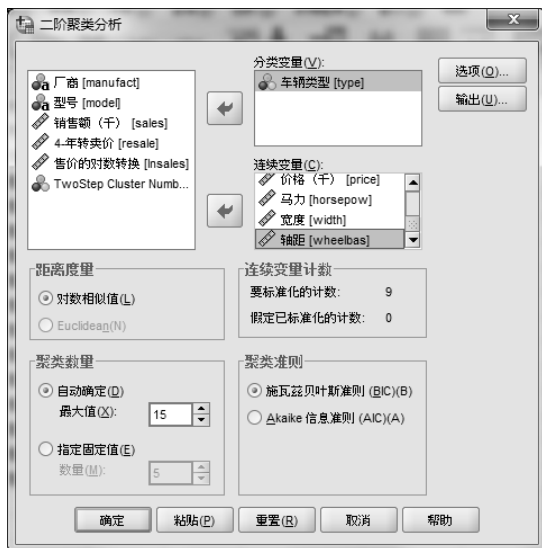


图 13-2 【二阶聚类分析】主对话框

的百分比,就被认为是叶子稀少。当把观测放到叶子稀少处,CF 树会长大。在树再次长大后,如果可能,待分类的观测会被放进 CF 树,否则,会被当作局外者丢弃。

如果不选择这个选项,聚类结束后,那些不能被指派到任何一类中的观测单独形成一类,称作局外类。

(2)【内存分配】栏。允许指定一个聚类过程中使用的最大存储空间(单位:MB)。如果两步聚类运行时需要占用的空间超出了该最大值,会使用磁盘存储内存中放不下的信息。默认的容量是 64MB。可以指定一个大于等于 64MB 的数值,或者请教系统管理员后再确定这个数值。这是处理和分析大样本所需要的。如果这个值太小,计算法则可能找不到正确的或者希望的类数。

(3)【连续变量的标准化】栏。聚类算法要求连续变量先完成标准化。任何连续变量都作为要被标准化的变量列在右侧框中。选择已经事先标准化的变量,单击向左箭头按钮,将其送入左侧的【假定已标准化的变量】框中。而要被标准化的变量留在右侧框中。可以对连续变量事先进行标准化,以便节省聚类过程所花费的计算时间,并简化操作。

(4)【高级】按钮。单击该按钮在【选项】可打开“高级选项”对话框,见图 13-4。

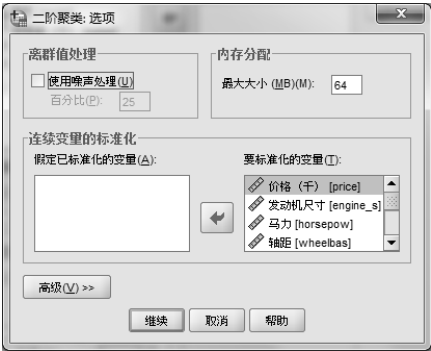


图 13-3 【二阶聚类: 选项】对话框

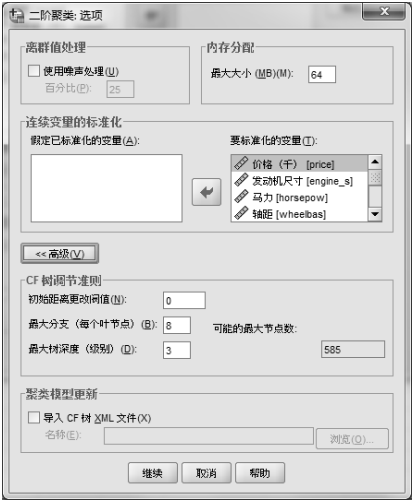


图 13-4 “高级选项”的【二阶聚类: 选项】对话框

①【CF 树调节准则】栏。聚类算法设置聚类特征(CF)树的特殊性,应该谨慎地改变有关的选项。

- 【初始距离更改阈值】。设置初始距离变化极限。这是用于增长 CF 树的起始极限。如果要把一个给定的观测插入到 CF 树的一个节点上,则产生的紧密性值应该比初始值要小,该叶子就不会被断开;如果密度值超过了初始值,则叶子会被断开,生成分支形成节点。系统默认值为 0,即开始时两个观测一定是各为一类。
- 【最大分支(每个叶节点)】。每个节点的最大分枝数。也就是一个节点可以具有的最大子节点数。系统默认值为 8。
- 【可能的最大节点数】。这是指可以由该分析过程产生的潜在的 CF 树节点的最大数,由公式  $(b^{d+1} - 1) / (b - 1)$  计算。这里的  $b$  是最大分枝数,  $d$  是最大树深度。根据系统默认值可以计算出默认的节点数为“585”。最低限度为每个节点需要 16B。一个很大的 CF 树会极大占用和消耗系统资源并反过来影响分析过程的执行。因此在这个对话框中要小心设置各参数。

- **【最大树深度】**。CF 树节点可以有的最大水平数。系统默认值为 3。

② **【聚类模型更新】** 栏。选择 **【导入 CF 树 XML 文件】** 将允许引入并用当前的数据文件修改以前生成的原聚类模型。引入的文件是 XML 格式的 CF 树。在主对话框中指定分析变量的顺序必须与以前分析时指定的变量顺序相同。除非明确地把新模型信息写到相同的文件名下，否则 XML 文件保持不变。

如果要更新模型，使用产生原模型时指定的与 CF 树有关的选项。更明确地说，使用生成原模型所用的距离测度、噪声处理、存储器设置或 CF 调谐判据等的设置，而不使用在当前对话框中的所选项和设置的参数。

**注意：**当对一个原聚类模型进行修改时，该过程假设在当前的工作数据文件中没有被选择的观测用于产生原分类模型。另外还假设，用于模型修改的观测与用于产生原模型的观测来自相同的总体。也就是说，两个数据集中的同名的连续变量的均值和标准差假设是相等的；分类变量的水平都相同。如果新的和旧的观测集来自不同的总体，为了得到最好的结果，应该根据两个数据集的组合来运行两步聚类分析。

#### 4. 输出选项

单击 **【输出】** 按钮，打开 **【两步聚类输出】** 对话框，如图 13-5 所示。

- (1) **【模型浏览器输出】** 栏。

① **【图表和表格】** 在输出浏览器中指定为评估字段的变量可以显示在模型浏览器作为聚类描述符。

模型浏览器中的表包括模型摘要和聚类-特征表。模型视图中的图形输出包括聚类质量图、聚成类的大小、变量重要性、聚类比较表和单元格信息。

② **【变量】** 栏里给出的是没有参与聚类分析的变量。可以选择并显示在模型浏览器中。

(2) **【工作数据文件】** 栏只有一个选项，在当前数据文件中产生一个新变量，即类成员变量。变量值为相应的观测属于哪一类。新变量名为 TSC<sub>*n*</sub>。*n* 为表明顺序的正整数，由系统自动给出。

- (3) **【XML 文件】** 栏。以 XML 格式输出文件。

① **【导出最终模型】**。把最终聚类模型输出到指定的文件。

② **【导出 CF 树】**。保存当前聚类树的状态到指定文件。该文件可以在以后用新数据分析的结果修改。

以上两项都要单击后面的 **【浏览】** 按钮，指定存储路径和文件名。

### 13.2.3 两步聚类分析实例

**【例 1】** 汽车制造商需要评价当前汽车市场，以确定车辆在市场上的竞争地位。通常可以对探访的数据进行分类来达到此目的，可以用自动的两步聚类分析来完成。

(1) 数据文件 data13-01 中包括了各种车辆发动机的构造、型号、价格和反映物理特性的数据。使用两步聚类分析过程根据价格和物理特性自动分类。变量名及其含义见表 13-1。

(2) 按 **【分析→分类→两步聚类】** 顺序单击菜单项，打开 **【二阶聚类分析】** 主对话框。

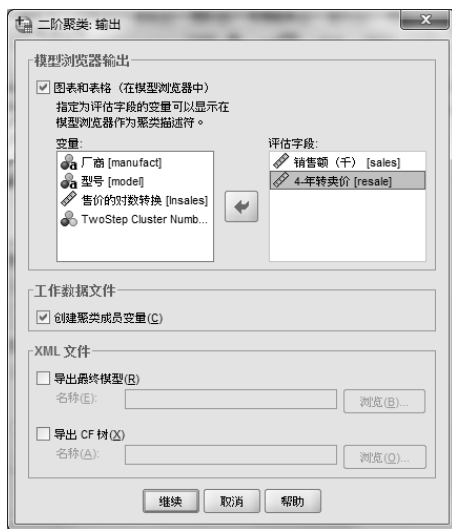


图 13-5 **【二阶聚类：输出】** 对话框

- ① 选择车的类型 type 变量，送入分类变量框中。
- ② 选择价格 price、发动机尺寸 engine\_s、马力 horsepower、轴距 wheelbas、宽度 width、长度 length、底盘重量 curb\_wgt、燃料容量 fuel\_cap、燃料功效 mpg 这 9 个连续型变量送入【连续变量】框中。

表 13-1 【例 1】变量说明

变量名	含义	变量名	含义	变量名	含义
manufact	厂商	price	价格	length	长度
model	型号	engine_s	发动机尺寸	curb_wgt	底盘重量
sales	销售量	horsepow	马力	fuel_cap	燃料容量
resale	4 年后销售量	wheelbas	轴距	mpg	燃料功效
type	类型	width	宽度		

③ 在主对话框中单击【输出】按钮，在【二阶聚类：输出】对话框的【模型浏览器输出】框中，选择唯一的选项[图表和表格(在输出浏览器中)指定为评估字段的变量可以显示在模型浏览器中作为聚类描述符]。

(3) 运行结果见表 13-2~表 13-5 和图 13-6~图 13-9。

图 13-6 所示是模型浏览器。

① 左侧窗口中的内容是在输出窗中的内容，双击输出窗中的输出，打开【模型浏览器】。利用左、右两侧窗口下边的【视图】下拉菜单观看更多的信息。

【模型概要】即聚类的综合信息。说明聚类算法是两步聚类；输入变量有 10 个；最后聚成 3 类。



图 13-6 模型浏览器

- ② 模型浏览器右侧窗口中有所聚成的类的大小饼图和需要的表格。
  - 饼图表明第一类观测数占 40.8%，第二类占 25.7%，第三类占 33.6%。
  - 表格中列出最小类和最大类中包含的观测数分别是 39 和 62，以及最大类与最小类观测数的比值 1.59。
- ③ 打开左侧窗口下面的【视图】下拉菜单，选择其中的【聚类】项，打开聚类表，如表 13-2



所示。表中的类自左至右是按类的大小排序的。第 1 行是所聚成的类号；第 2 行是可以由用户自己添加标签的单元格。双击单元格即可以添加文字；第 3 行是可以由用户自己对各类添加说明的单元格；第 4 行是各类的大小，即各类中的观测数；第 5 行是作为输入变量的分类变量，以下 9 行是 9 个连续型变量。各单元格中列出各类、各变量的均值。表格中的小方框是当鼠标指向一个单元格时，显示的该类该变量的信息，包括变量名、该变量在聚类过程中的重要性、均值。各列就是各类的类中心，由 9 个变量的均值组成。

它表明了连续型变量很好地把各类分开了。1 类中的车辆便宜、小(长度、宽度都小)，燃料功效最高；2 类中的车辆特征是价格适度、汽缸较大；3 类中的车辆昂贵、大，燃烧效率适度。

(4) 打开模型浏览器右侧窗口下面的【视图】菜单，可选择预测变量的重要性项。右侧窗口显示重要性图，见图 13-7。重要性图以条形图表示，最下面的标尺表明条形图越短，重要性越差。重要性最高的是类型 type，重要性最差的是价格 price。

表 13-2 模型浏览器中的聚类表

输入（预测变量）重要性  
■ 1.0 ■ 0.8 ■ 0.6 ■ 0.4 ■ 0.2 □ 0.0

聚类	1	3	2
标签			
说明			
大小	<div><div></div></div> 40.8% (62)	<div><div></div></div> 33.6% (51)	<div><div></div></div> 25.7% (39)
输入	车辆类型 Automobile (98.4%)	车辆类型 Automobile (100.0%)	车辆类型 Truck (100.0%)
	长度 178.24	长度 194.69	长度 191.11
	底盘重量 2.84	底盘重量 3.58	底盘重量 3.97
	燃料容量 14.98	燃料容量 18.44	燃料容量 22.06
	燃料效率 27.24	燃料效率 23.02	燃料效率 19.51
	发动机尺寸 2.19	发动机尺寸 3.70	发动机尺寸 3.56
	价格（千） 19.62	价格（千） 37.30	价格（千） 26.56
	马力 143.24	马力 232.9 马力重要性 = 0.63 均值: 232.96	马力 187.92
	宽度 68.54	宽度 72.92	宽度 72.74
	轴距 102.60	轴距 109.02	轴距 112.97

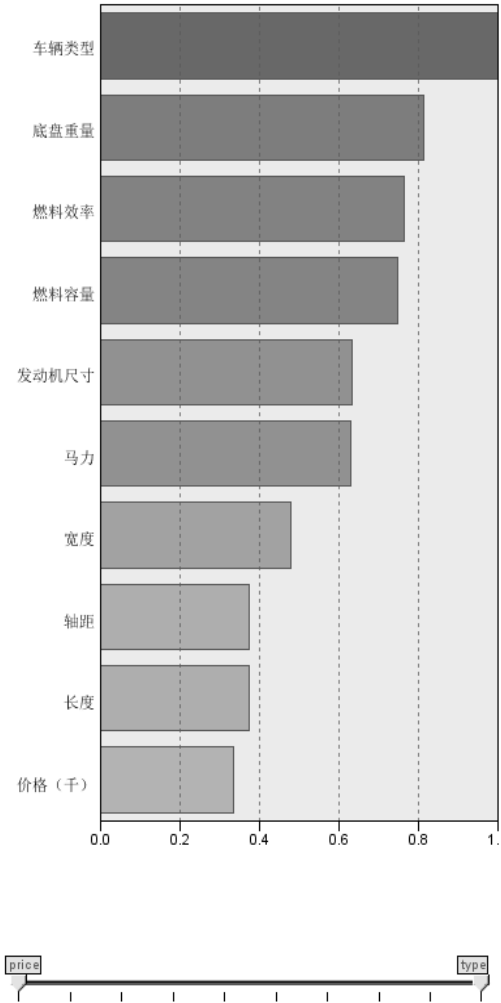


图 13-7 预测变量重要性图

使用两步聚类过程 Two Step Cluster Analysis 已经把车辆分为明显的 3 类。为了更好地在内部将各类分开, 还需要收集有关车辆的其他方面的信息, 如碰撞试验的成绩或有用的其他项目的信息。

## 13.3 快速聚类

### 13.3.1 快速聚类概述

当要聚成的类数确定时, 使用快速聚类过程可以很快将观测分到各类中去。其特点是处理速度快、占用内存少, 适用于大样本的聚类分析。

K-均值聚类执行快速聚类命令, 使用  $k$  均值分类法对观测进行聚类, 可以完全使用系统默认值执行该命令, 也可以对聚类过程设置各种参数进行人为的干预。例如, 可以事先指定把数据文件中的观测分为几类, 指定使聚类过程中止的收敛判据或迭代次数, 指定聚类结束后在的输出窗口中显示哪些内容, 是否将聚类结果或中间数据存入输出数据文件, 指定其文件名以及把哪些数据存入数据文件等。

进行快速聚类首先要选择用于聚类分析的变量和类数。参与聚类分析的变量必须是数值型变量, 且至少要有一个。为了清楚地表明各观测最后聚到哪一类, 还应该定一个表明观测特征的变量作为标识变量, 如“编号”、“姓名”之类的变量。聚类数必须大于等于 2, 但类数不能大于数据文件中的观测数。

如果选择了  $n$  个数值型变量参与聚类分析, 最后要求聚类数为  $k$ , 那么可以由系统首先选择  $k$  个观测(也可以由读者指定)作为聚类的种子,  $n$  个变量组成  $n$  维空间。每个观测在  $n$  维空间中是一个点。 $k$  个事先选定的观测就是  $k$  个聚类中心点, 也称为初始类中心。按照离这几个类中心的距离最小原则, 把观测分派到各类中心所在的类中去, 构成第一次迭代形成的  $k$  类。根据组成每一类的观测, 计算各变量均值。每一类中的  $n$  个均值在  $n$  维空间中又形成  $k$  个点。这就是第二次迭代的类中心。按照这种方法依次迭代下去, 直到达到指定的迭代次数或达到中止迭代的判据要求时, 迭代停止, 聚类过程结束。

快速聚类使用的是欧氏距离平方, 各变量权数相等。如果使用其他统计量进行聚类, 必须使用系统聚类法进行聚类分析。快速聚类变量必须是连续变量。如果测定变量值的单位不同, 应对聚类变量使用描述性统计分析过程进行标准化后再进行聚类分析, 否则会得出错误的结论。如果聚类变量是计数变量或二值变量, 则使用系统聚类分析过程进行聚类分析。

### 13.3.2 快速聚类过程

快速聚类过程适用于对大样本进行快速聚类, 尤其是对形成的类的特征(各变量值范围)有了一定认识时, 此方法使用起来会更加得心应手。操作方法分为以下几步:

(1) 建立或读入数据文件后, 按【分析→分类→K-均值聚类】顺序单击菜单项, 打开如图 13-8 所示的对话框。

(2) 指定分析变量和标识变量。在源变量表中选择参与聚类分析的数值型变量, 移到右侧的【变量】框中; 选择能唯一标识各观测的变量, 送入【个案标记依据】框中。

(3) 确定分类数。在【聚类数】框中显示系统默认分为两类。可按分析要求输入分类数。

(4) 选择聚类方法。在【方法】栏中选择一种聚类方法。

①【迭代与分类】。聚类的迭代过程中使用 K-均值算法不断计算类中心, 并根据结果更换

类中心，把观测分派到与之最近的一类中心为标志的类中去。这是系统默认选项。

②【仅分类】。根据初始类中心进行聚类。在聚类过程中不改变类中心。

(5) 在【聚类中心】栏内选择初始类中心。

① 读取初始聚类中心。要求使用指定数据文件中的观测作为初始类中心。选择此项后，还需要选择：

- 【打开数据集】。如果包含种子观测的数据文件已经打开，在下拉列表选择一个其观测作为初始类中心的数据集。
- 【外部数据文件】。如果包含种子观测的数据文件没有打开，单击【文件】按钮，在【读取文件】对话框中指定文件所在位置(路径)和文件名。该文件的观测作为初始类中心的数据，单击【打开】按钮返回。在外部数据文件下，【文件】按钮后面显示包括路径的文件全名。

选择读取初始类中心，需要事先建立一个数据集，其中观测的数目与要聚成的类数相等，每个观测都由参与聚类的变量值组成。

②【写入最终聚类中心】。要求把聚类结果中的各类中心数据保存到指定的文件中，该文件可以作为以后聚类的初始类中心文件。

- 【新数据集】。在后框中输入数据文件名，运行结果会把最后结果的类中心保存在指定的文件中。注意这个文件无须指定保存位置，所以结束 SPSS 前要保存这个文件，否则会丢失。
- 【数据文件】。单击【文件】按钮，在【写入文件】对话框中指定文件保存位置(路径)和文件名，单击【保存】按钮返回。在【文件】按钮后面显示包括路径的文件全名。

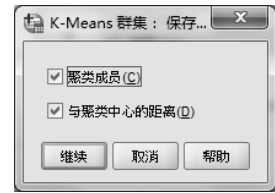


图 13-9 【K-Means 群集：保存新变量】对话框

(7) 控制聚类分析过程的选项。

单击【迭代】按钮，打开【K-均值聚类分析：写入文件】对话框，见图 13-10。只有在主对话框【方法】栏中选择了【迭代与分类】，才会激活此项，打开此对话框对迭代次数和聚类判据进行进一步选择。

①【最大迭代次数】框。限定 K-均值算法中的迭代次数。当达到限定的迭代次数时，即使没有满足收敛判据，迭代也停止。系统默认值为 10，选择范围为 1~999。

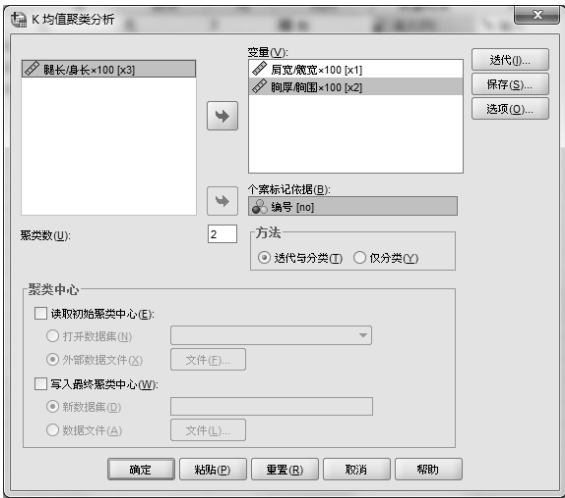


图 13-8 【K 均值聚类分析】主对话框

②【收敛性标准参数】框。指定 K-均值算法中的收敛判据，其值必须大于等于 0，且小于 1，默认值为 0。该项数值等于  $N\%$  的含义是：当两次迭代计算的最小的类中心的变化距离小于初始类中心距离的  $N\% \times 100$  时迭代停止。例如判据设置为 0.02，当一次完整的迭代不能使任何一个类中心距离的移动(变化量)与原始类中心距离的比小于 2%时，迭代停止。

③ 若设置了以上两个参数，在迭代过程中，满足了其中一个参数，迭代就停止。

④【使用运行均值】。限定在每个观测被分配到一类后即刻计算新的类中心。如果不选择此项，在完成所有观测的一次分配后再计算各类的类中心，这样节省迭代时间。

(8) 单击【选项】按钮，打开如图 13-11 所示的【K 均值聚类分析：选项】对话框，指定要计算的统计量和对带有缺失值的观测的处理方式。

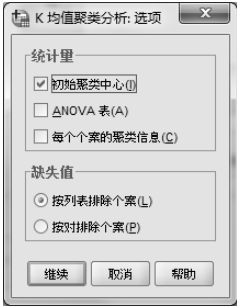
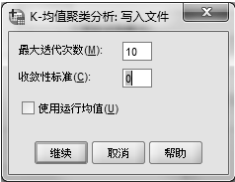


图 13-10 【K-均值聚类分析：写入文件】对话框

图 13-11 【K 均值聚类分析：选项】对话框

- ①【统计量】栏。选择要计算和输出的统计量。
- 【初始聚类中心】。
  - 【ANOVA 表】。方差分析表。
  - 【每个个案的聚类信息】。如最终所属类和该观测距所属类中心的距离。
- ②【缺失值】栏。选择一种处理带有缺失值观测的方法。
- 【按列表排除个案】。从分析中剔除在变量表中的变量带有缺失值的观测。
  - 【按对排除个案】。只有当一个观测的全部聚类变量值均缺失时才将其从分析中剔除，否则根据所有其他非缺失变量值，把它分配到最近的一类中去。

13.3.3 快速聚类分析实例

【例 2】 本例对游泳运动员进行聚类，以便分项。为简化问题，仅以 10 名运动员的 3 项测试数据为例。其中变量为 x1(肩宽/髋宽×100)、x2(胸厚/胸围×100)、x3(腿长/身高×100)，预计按姿势分为蝶泳、仰泳、蛙泳、自由泳 4 类。打开数据文件 data13-02。

- (1) 操作步骤。
- ① 按【分析→分类→K-均值聚类】顺序单击菜单项，打开主对话框。
- ② 本例要求根据 x1~x3 进行聚类，因此选择这 3 个变量，移至【变量】框中。选择变量 no 作为标识变量送入【个案标记依据】框中。
- ③ 泳姿分 4 类，在【聚类数】框中输入“4”。
- 其余使用系统默认值。提交系统执行。
- (2) 输出结果(见表 13-3~表 13-6)。
- 表 13-3 所示初始类中心。由于没有指定聚类的初始类中心，此表中的作为类中心的观测是由系统确定的。表中给出作为 4 类初始类中心的观测各变量值。

表 13-4 所示是两次迭代后类中心的变化。由于没有指定迭代次数或收敛判据，因此使用系统默认值：最大迭代次数为 10，收敛判据为 0。本快速聚类过程执行两次迭代后，类中心的变化为 0，迭代就停止了。表 13-4 给出了每次迭代类中心的变化量。

表 13-3 初始类中心

	聚类			
	1	2	3	4
肩宽/腕宽×100	125	122	120	120
胸厚/胸围×100	20	18	17	19
腿长/身长×100	44	43	42	44

表 13-4 两次迭代后类中心的变化

迭代	聚类中心内的更改			
	1	2	3	4
1	.707	.354	.707	.707
2	.000	.000	.000	.000

a. 由于聚类中心内没有改动或改动较小而达到收敛。任何中心的最大绝对坐标更改为 .000。当前迭代为 2。初始中心间的最小距离为 2.449。

表 13-5 给出了聚类结果形成的 4 类的类中心的 3 个变量的值。右表显示的是聚类结果，每类中观测的数目。除第二类有 4 个外，其余各类均有 2 名运动员。

表 13-5 最终的四类的类中心

	聚类			
	1	2	3	4
肩宽/腕宽×100	125	122	121	121
胸厚/胸围×100	20	18	17	19
腿长/身长×100	45	43	42	45

表 13-6 和聚类总结

聚类	1	2.000
	2	4.000
	3	2.000
	4	2.000
有效		10.000
缺失		.000

由上述结果输出可以看出，采用系统默认值的输出结果并不令人满意。若想知道某个观测属于哪一类，从输出信息中找不到。因此需要使用选项。

【例 3】 指定初始类中心的聚类方法例题。

仍使用数据文件 data13-02。已知 no=9、8、4、6 的 4 名运动员分别是蝶、仰、蛙、自由 4 种泳姿成绩的突出者，以这 4 个观测作为初始聚类中心进行聚类。操作步骤如下：

(1) 建立包含初始聚类中心 4 个观测的数据文件，类中心数据文件(也称种子数据文件) data13-02a 存入磁盘中。要求种子数据文件：

- ① 其格式必须与中数据文件 data03-02 的格式相同。
- ② 文件中的变量必须在当前工作数据文件中存在，并且变量名相同，在即将进行的快速聚类中也选择相同的变量作为聚类变量。
- ③ 其中的观测数必须与在主对话框中指定的类数相同。
- ④ 有一个表明类号的变量，变量名为 cluster\_。

该数据文件可以是前一次快速聚类产生的输出文件，也可以根据经验找出的最具代表性的观测作为初始类中心，或称聚类的种子。

(2) 将已经存在的原始数据文件 data13-02 调入，显示在当前数据窗口中。打开作为种子的数据文件 data13-02a。

(3) 首先，按【例 2】中的(1)~(3)步选择聚类变量、标识变量，指定分类数。

(4) 在【聚类中心】栏内选中【读取初始聚类中心】，选择【打开数据集】，在其下拉列表中显示 data13-02a.sav[datasetn]，即已经打开的数据文件名指定为初始类中心文件。

(5) 选中【写入最终聚类中心】，选择【数据文件】选项，保存聚类结果的类中心数据为数据文件。单击【文件】按钮，指定存储作为以后种子观测的数据文件。指定存取路径和文件名(data13-02b)。

- (6) 选择聚类方法。在主对话框【方法】栏中选择【迭代与分类】。
- (7) 聚类过程控制参数仍选用系统默认值。【迭代】项内的值保持不变。
- (8) 单击【保存】按钮，在对话框中选择【聚类成员】和【与聚类中心的距离】两个复选项。
- (9) 单击【选项】按钮，打开相应的对话框，选中【统计量】栏中的全部复选项。由于数据文件中没有缺失值，故保持【缺失值】栏中系统默认的处理方式。
- (10) 在主对话框中单击【确定】按钮提交系统执行。
- (11) 输出结果见表 13-7~表 13-13、图 13-13 和图 13-14。数据文件中的聚类种子数据见图 13-12。
- (12) 结果解释。

表 13-7 中的初始类中心是指定的种子文件 data13-02a 中的数据。

表 13-8 表明共经过两次迭代完成聚类。第一次迭代 1~4 类的类中心与初始类中心之间的距离分别为 0.707、0.707、0.745、1.054。从操作过程看到结束聚类过程的判据有两个，一个是最大迭代次数为 10，另一个是类中心变化距离为 0。从表 13-10 看到，当进行第二次迭代后，类中心几乎没有变化，使用判据 0，结束了聚类过程。

表 13-7 初始类中心

	聚类			
	1	2	3	4
肩宽/臂宽×100	124	120	122	122
胸厚/胸围×100	20	19	19	17
腿长/身长×100	45	44	43	42

从 FILE 子命令中输入

表 13-8 迭代过程中类中心的变化量

迭代	聚类中心内的更改			
	1	2	3	4
1	.707	.707	.745	1.054
2	.000	.000	.000	.000

a. 由于聚类中心内没有改动或改动较小而达到收敛。任何中心的最大绝对坐标更改为 .000。当前迭代为 2。初始中心间的最小距离为 2.236。

表 13-9 给出了聚类结果，即每个观测用 no 标识，表头的“编号”为变量 no 的标签。聚类列的值为类号，表明各观测最终被分配到哪一类，距离的值为该观测在三维坐标中的点与类中心点的距离。如果选择的类中心是各类最具代表性的观测，则距离值越大，与该类代表性观测的差异越大。

表 13-10 给出了 4 个类中心的 3 个变量值，即类中心在三维坐标空间中的位置。

表 13-11 给出的是聚类结束时，两两类中心间的距离。表格第 1 行和左侧第 1 列均为类号。两类间的距离在行、列交叉点单元格中。

表 13-9 各观测所属类成员表

案例号	编号	聚类	距离
1	1	1	.707
2	2	3	.745
3	3	4	1.054
4	4	1	.707
5	5	3	.471
6	6	2	.707
7	7	4	.667
8	8	3	.745
9	9	4	1.054
10	10	2	.707



图 13-12 聚类种子数据

表 13-12 所示是方差分析表。3 个变量中任意一个变量的类间均方值(Cluster MS)都远远大于类内的误差均方值(Error MS)。从概率值来看，3 个变量使类间无差异的假设检验的概率均小于 0.1%。方

差分析结果表明，参与聚类分析的 3 个变量能很好地区分各类，类间的差异足够大。聚类的方差分析检验的零假设应该是：类均值相等(各类间无差异)。该分析结果可用于描述分类的目的。

表 13-13 所示是聚类总结，给出了各类的观测数、参与分析的合法观测数有效的和缺失值数。可以看出，指定了初始类中心(种子)的结果与没使用初始类中心的结果略有不同(与表 13-6 比较)。

表 13-10 最终的类中心的变量值

	聚类			
	1	2	3	4
肩宽/腕宽×100	125	121	122	121
胸厚/胸围×100	20	19	18	17
腿长/身长×100	45	45	43	42

表 13-11 最终的类中心间的距离

聚类	1	2	3	4
1		4.123	3.613	5.411
2	4.123		2.014	3.504
3	3.613	2.014		2.000
4	5.411	3.504	2.000	

表 13-12 方差分析

	聚类		误差		F	Sig.
	均方	df	均方	df		
肩宽/腕宽×100	6.644	3	.611	6	10.873	.008
胸厚/胸围×100	3.911	3	.111	6	35.200	.000
腿长/身长×100	4.644	3	.278	6	16.720	.003

表 13-13 聚类总结

聚类	1	2.000
	2	2.000
	3	3.000
	4	3.000
有效		10.000
缺失		.000

F 检验应仅用于描述性目的，因为选中的聚类将被用来最大化不同聚类中的案例间的差别。观测到的显著性水平并未据此进行更正，因此无法将其解释为是对聚类均值相等这一假设的检验。

图 13-13 所示是当前工作数据文件。根据指定的选项共建立了 2 个新变量，显示在工作数据文件窗口中。QCL1 是类号，QCL2 是观测距所属类的类中心之间的距离。

输出数据文件中只有最终的类中心数据，见图 13-14。此输出数据文件可以作为对另一个样本进行快速聚类的初始类中心。

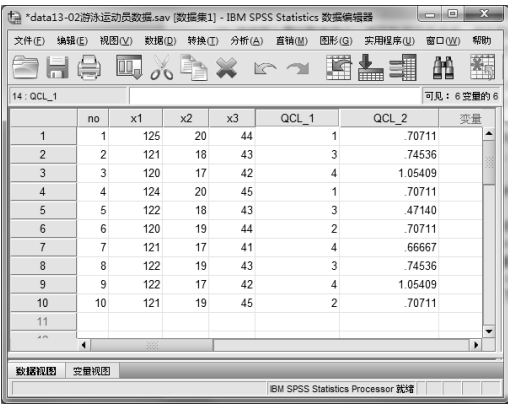


图 13-13 工作数据文件中的新变量

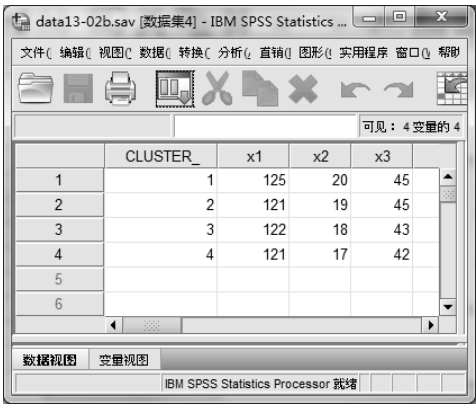


图 13-14 输出数据文件

## 13.4 系统聚类

### 13.4.1 系统聚类概述

#### 1. 系统聚类的概念

聚类的方法有多种，除了前面介绍的两步聚类和快速聚类外，最常用的是系统聚类。根据聚类过程不同，系统聚类又分为分解法和凝聚法。

(1) 分解法。聚类开始时把所有个体(观测或变量)都视为属于一大类，然后根据距离和相似性逐层分解，直到参与聚类的每个个体自成一类为止。

(2) 凝聚法。聚类开始时把参与聚类的每个个体(观测或变量)视为一类,根据两类之间的距离或相似性逐步合并,直到合并为一个大类为止。

无论哪种方法,其聚类原则都是相近的聚为一类,即距离最近或最相似的聚为一类。实际上,以上两种方法是方向相反的两种聚类过程。

## 2. 系统聚类过程的功能

系统聚类的方法包括样品聚类(Q 型)和变量聚类(R 型)。通常情况下在聚类进行之前,应该先根据反映各类特性的变量对原始数据进行标准化处理,即利用标准化方法对原始数据进行一次转换,并计算相似性测度或距离测度,然后系统聚类过程根据转换后的数据进行聚类分析。SPSS 的系统聚类的各方法都包含了邻近度过程对数据的处理,系统聚类过程对数据的聚类。给出的统计量可以帮助用户确定最好的分类结果。

系统聚类过程可以通过 Plot 选项给出两种图: Dendrogram(树形图)和 Icicle(冰柱图)。

系统聚类过程的输出项可以选择,还可以建立新变量,把聚类结果,即每个个体被分派到的类编号作为新变量的值,保存到当前的工作数据文件中。

## 3. 在系统聚类过程中使用的术语

(1) 聚类方法。实现系统聚类的具体方法有许多种,各种方法的区别在于如何定义和计算两项(两个个体、两类或个体与类)之间的距离或相似性。这一点体现在聚类方法的一系列选项上。如果不熟悉对聚类方法的定义,可以使用系统默认的方法。需要确定的选项有:

- 聚类法的选择。定义计算两项间距离和相似性的方法,默认使用组间平均连接法。
- 测度方法的选择。对距离和相似性的测度方法又有多种,这一点体现在测度方法的选择上。如果对测度方法不熟悉,可以采用系统默认的欧氏距离平方。

定义距离和相似性的方法不同,测度距离和相似性的算法就不同,会导致聚类结果稍有区别,但大体上是一致的。

(2) 标准化。如果参与聚类的变量的量纲不同会导致错误的聚类结果。因此在聚类过程进行之前必须对变量值进行标准化,即消除量纲的影响。用不同方法进行标准化,聚类结果也会有所不同,因此在选择标准化方法时要注意变量的分布。如果是正态分布应该采用 Z 分数法。如果参与聚类的变量量纲相同,可以选择对数据不进行标准化处理。

(3) 树形图。表明每一步中被合并的类及系数值,把各类间的距离转换成 1~25 间的数值。

(4) 冰柱图。把聚类信息综合到一张图上。如果作纵向冰柱图,则参与聚类的个体各占一行,标以个体(观测或变量)号或在图纸允许的情况下标以个体的标签,聚类过程中的每一步占一行,标以步的顺序号;如果作横向冰柱图,则参与聚类的个体(观测或变量)各占一行,聚类的每一步占一列。如果不指定加以限制的选项,则显示聚类的全过程。

树形图和冰柱图都是最后确定分类结果的重要手段。因为无论凝聚法还是分解法均不给出确定的分类结果,最后的分类结果需要用户根据研究的对象和研究目的自己确定。

### 13.4.2 系统聚类过程

无论是观测聚类还是变量聚类均按下述步骤进行。在叙述操作步骤的过程中对涉及的选项及其含义分别加以说明。

(1) 在数据窗口中建立工作数据文件。



(2) 按【分析→分类→系统聚类】顺序单击菜单项，打开如图 13-15 所示的【系统聚类分析】主对话框。

(3) 在对话框中部的【聚类】栏中选择聚类类型。

- ①【个案】。要进行观测(Q 型)聚类。
  - ②【变量】。要进行变量(R 型)聚类。
- (4) 指定参与分析的变量，即能反映分类特征的变量，送入【变量】框中。如果作观测(样品)聚类，还要选择能唯一标识观测的变量，移到右侧的【标注个案】框中。



图 13-15 【系统聚类分析】主对话框

(5) 如果参与分析的变量量纲一致，则不必对数据进行标准化，其余选项全部选择系统默认值，此时就可以单击【确定】按钮提交执行了。

(6) 确定聚类方法。

在主对话框中，单击【方法】按钮，打开如图 13-16 所示的【系统聚类分析：方法】对话框。根据需要指定聚类方法、距离测度方法、对数值进行转换(标准化)的方法。距离测度、对数据继续转换的方法算法请参考本书附录 A。



图 13-16 【系统聚类分析：方法】对话框

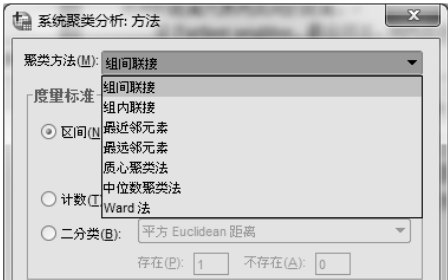


图 13-17 【聚类方法】下拉列表

- 【聚类方法】选择。单击【聚类方法】下拉列表，打开如图 13-17 所示的聚类方法清单。
- 【组间联接】。合并两类的结果使所有的两两项对之间的平均距离最小。项对的两个成员分别属于不同的类。该方法中使用的是各对之间的距离，既非最大距离，也非最小距离。
- 【组内联接】。合并为一类后，类中的所有项之间的平均距离最小。该距离就是合并后的类中所有可能的观测对之间的距离平方。
- 【最近邻元素】。合并最近的或最相似的两项，用两类间最近点间的距离代表两类间的距离。
- 【最远邻元素】。用两类间最远点的距离代表两类间的距离，也称为完全连接法。
- 【质心聚类法】。像计算所有各项均值之间的距离那样计算两类之间的距离，该距离随聚类的进行不断减小。
- 【中位数聚类法】。以各类中的变量值中位数为类中心。

- **【Ward 法】**。Ward 最小方差法。以方差最小为聚类原则。
- ② 对距离和相似性测度方法的选择。在**【数量标准测度】**栏中指定用哪两点间的距离作为确定是否合并的距离。距离的具体计算方法还根据参与距离计算的变量类型，从以下 3 种菜单选择其一，打开选择菜单后再进行具体方法的选择。这 3 个菜单分别对应于等间隔测度的变量(一般为连续变量)、计数变量(一般为离散变量)和二值变量。
  - 对于等间隔测度的变量，可在**【系统聚类分析：方法】**对话框**【度量标准】**栏**【区间】**下拉列表中选择，见图 13-18。这些方法是：欧几里得距离(又称欧氏距离)、欧氏距离平方、余弦相似性测度、Pearson 皮尔逊相关、Chebychev 切贝谢夫距离、Block 块布洛克距离、Minkowski(明可斯基)距离。选择 Minkowski 后，还须输入乘方次数和开方次数  $p$ 。设定距离选项允许用户自定义距离计算方法，两项之间的距离用各项值之间差值的  $p$  次幂绝对值之和的  $r$  次方根表示，选择此项后，还须输入乘方次数  $p$  和开方次数  $r$ 。以上各选项的计算方法参见附录 A 有关内容。
  - 对于计数变量(离散型变量)，选择**【计数】**项。可在下拉列表中选择不相似性测度的方法： $\chi^2$  卡方测度、Phi 方( $\Phi^2$ )测度。各选项的计算方法请见附录有关内容。
  - 对于二值变量，选择**【二分类】**项，在如图 13-19 所示的下拉列表中选择距离或不相似性测度的方法。各方法的计算公式见附录有关内容。



图 13-18 **【系统聚类分析：方法】**对话框  
**【度量标准】**栏**【区间】**下拉列表

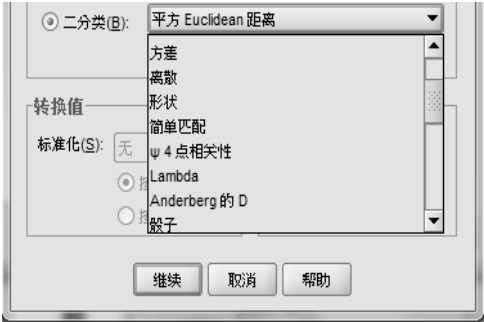


图 13-19 **【系统聚类分析：方法】**对话框  
**【度量标准】**栏**【二分类】**下拉列表

首先应该明确，对二值变量，系统默认用 1 表示某特性的出现(或发生、存在等)，用 0 表示某特性不出现(或不发生、不存在)。

在使用程序语句进行计算时如果指定两个参数，系统认为第一个参数表示某事件发生，第二个参数表示某事件不发生；如果只指定一个参数，系统认为该参数表示事件发生，其他值表示事件不发生。选项共 27 项。有关的约定和计算方法及解释参见附录 A 的有关内容。

这 27 项测度二分类变量类间距离的方法有：Euclidean 距离、Euclidean 距离平方、尺度差分、模式差分、方差、离散、形状、简单匹配、 $\Psi 4$  点相关性、Lambda、Anderberg 的 D、骰子、Hammann、Jaccard、Kulczynski1、Kulczynski2、Lance 和 Williams、chiai、Rogers 和 Tanimoto、Russel 和 Rao、Sokai 和 Sneath 1、Sokai 和 Sneath 2、Sokai 和 Sneath 3、Sokai 和 Sneath4、Sokai 和 Sneath5、Yule 的 Y、Yule 的 Q。

从下拉列表中选择一种测度方法。还可以改变表示某事件存在于不存在(发生与不发生或说某特性出现与不出现的值)，在**【存在】**和**【不存在】**框中输入用户自己定义的值(当然应该与数据文件中有关的二元变量的值一致)，定义后，系统将忽略其他值。如果不进行自定义，那么，1 代表某事件存在，0 代表某事件不存在。

③【转换值】栏。如果在【度量标准】栏中选择的是间隔变量的度量，则【转换值】栏被激活，应该在【转换值】栏的【标准化】下拉列表中选择标准化的方法，如图 13-20 所示。只有等间隔测度的数据(选择了【区间】)或计数数据(选择了【计数】)才可以进行标准化。对数据进行标准化的方法有：

- 【无】。不进行标准化，是系统默认值。
- 【Z 得分】。把数值标准化到 Z 分数。
- 【全距从-1 到 1】。把数值标准化到-1~1 范围内。
- 【全距从 0 到 1】。把数值标准化到 0~1 的范围内。
- 【1 的最大量】。把数值标准化到最大值为 1。
- 【均值为 1】。把数值标准化到均值为 1。
- 标准差为 1。把数值标准化到单位标准差。

有关标准化方法的具体算法参见附录 A 的有关内容。

④ 测度的转换方法选择。可选择的转换方法在【转换度量】栏中。

- 【绝对值】。对距离值取绝对值。当数值符号表示相关方向且只对负相关关系感兴趣时使用此方法进行变换。
- 【更改符号】。把相似性值变为不相似性值或相反，用求逆的方法使距离顺序颠倒。
- 【重新标度到 0-1 全距】。通过先减去最小值然后除以全距(两极差)的方法使距离标准化。

对已经按某种计算方法计算了相似性或不相似性测度的一般不再使用此方法进行转换。如果使用的是已经存在的矩阵，可以选择此类选项，对输入矩阵进行必要的转换。

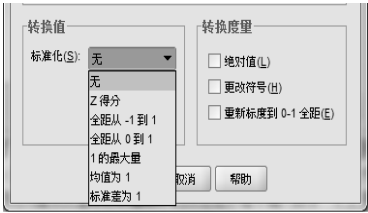


图 13-20 【转换值】栏【标准化】下拉列表

步骤(6)的 4 组选项选择完成后，单击【继续】按钮，返回主对话框。

(7) 选择要求输出的统计量。在主对话框中单击【统计量】按钮，打开相应的对话框，见图 13-21，指定要输出的统计量。

①【合并进程表】。要求作凝聚状态表。凝聚状态表显示聚类过程中每一步合并的两项(观测与观测、观测与类、类与类)、被合并的两项之间的距离以及观测或变量加入到一类的类水平，因此可以根据此表跟踪聚类的合并过程。由于最接近的两类先聚为一类，因此可以通过聚类过程仔细地查看哪些观测更接近一些。



图 13-21 【系统聚类分析: 统计量】对话框

②【相似性矩阵】。要求输出各项间的距离矩阵。以矩阵形式给出各项之间的距离或相似性测度值。产生什么类型的矩阵(相似性矩阵或不相似性矩阵)取决于在【系统聚类分析: 方法】对话框中【度量标注】栏中的选择。

注意：如果项数很大(观测数或变量数很大)则该选项产生的输出量也会很大。

③【聚类成员】栏。要求显示每个观测被分派到的类(即分类结果，各观测属于哪一类)或显示若干步凝聚过程。可以用下面的选项进一步选择：

- 【无】。不显示类成员表。是系统默认选项。
- 【单一方案】。要求聚为指定类数时，列出各观测所属的类。在【聚类数】框中输入限定显示的类数，该数值必须是大于 1 且小于等于参与聚类的观测或变量总数的整数。例如，在矩形

框中输入数字“3”，则会在输出窗口中显示聚为 3 类时每个观测属于 3 类中的哪一类。

● **【方案范围】**。要求列出某个范围中每一步聚类过程和各观测所属的类。

A. 在**【最小聚类数】**框中输入一个最小类数值。

B. 在**【最大聚类数】**框中输入最大类数值。

这两个数值必须是不等于 1 的正整数，最大类数值不能大于参与聚类的观测数或变量总数，例如，选择此选项并在上、下两个矩形框中分别输入了“3”和“5”，将在输出窗中显示 3 个结果：观测(或变量)被聚为 3 类、4 类、5 类时各观测(或变量)被分派到哪一类。

以上内容所涉及的变量或观测取决于在主对话框中的**【聚类】**栏中选择的是**【个案】**还是**【变量】**。

(8) 选择统计图表。在主对话框中，单击**【绘制】**按钮，打开如图 13-22 所示**【系统聚类分析：图】**对话框。

① **【树状图】**选项。

② **【冰柱图】**栏。对于生成什么样的冰柱图还可以进一步用以下选项确定：

● **【所有聚类】**。聚类的每一步都表现在图中。可用此种图查看聚类的全过程，但如果参与聚类的个体很多会造成图过大，没有必要。

● **【聚类的指定全距】**。指定显示的聚类范围。当选择此项时，在**【开始聚类】**框中输入要求显示聚类过程的起始步数，在**【停止聚类】**框中输入中止于哪一步，在**【排序标准(输出间隔)】**框中输入两步之间的增量。输入的数字必须是正整数。例如，输入的数字是：**【开始聚类】**“3”，**【停止聚类】**“10”，**【排序标准】**“2”。生成的冰柱图从第三步开始，显示第三、五、七、九步聚类的情况。

● **【无】**。不生成冰柱图。

③ 显示方向可以在**【方向】**栏中确定。

● **【垂直】**。显示纵向的冰柱图。

● **【水平】**。显示水平的冰柱图。

(9) 生成新变量的选项。聚类分析的结果可以用新变量保存在工作数据文件中。单击主对话框的**【保存】**按钮，打开如图 13-23 所示的对话框。可以看出，只能生成一个表明参与聚类的个体最终被分配到哪一类的新变量。在对话框**【聚类成员】**栏中可以选择是否建立新变量和所建新变量的含义。

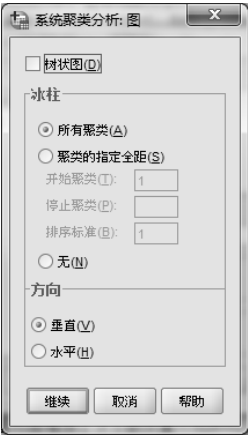


图 13-22 **【系统聚类分析：图】**对话框



图 13-23 **【系统聚类分析：保存】**对话框

① **【无】**。不建立新变量。

② **【单一方案】**。即单一结果，生成一个新变量，表明每个个体聚类最后所属的类。在**【聚类数】**框中指定类数，如果输入“5”，则新变量值的范围为 1~5。

③ **【方案范围】**。即新变量表示指定范围内的结果。生成若干个新变量，表明聚为若干个类时，每个个体聚类后所属的类。把表示从第几类显示到第几类的数字分别输入到**【最小聚类数】**框和**【最大聚类数】**框中。例如，输入的数值是：**【最小聚类数】**“4”，**【最大聚类数】**“6”，在聚类结束后在数据窗口中原变量后面增加了 3 个新变量，分别表示分为 4 类时、分为 5 类和分为 6 类时的聚类结果，即各观测分别属于哪一类。

### 13.4.3 样品系统聚类分析实例

**【例 4】** 数据文件 data13-03 为一组有关 12 盎司啤酒成分和价格的数据，变量包括啤酒品牌(beername)、热量卡路里(calorie)、钠含量(sodium)、酒精含量(alccohol)、价格(cost)。要求根据 12 盎司啤酒的各成分含量及 12 盎司啤酒价格对 20 种啤酒进行分类。

(1) 读取数据文件后，操作步骤如下。

① 按**【分析→分类→系统聚类】**顺序单击菜单项，打开**【系统聚类分析】**主对话框。

② 选择热量 calorie、钠含量 sodium、酒精含量 alccohol、价格 cost 这 4 个变量为分析变量，移到**【变量】**框中。选择啤酒品牌 beername 作为标识变量，移到**【标注个案】**框中。

③ 选择 Q 型聚类。在**【聚类】**栏中选择系统默认的**【个案】**聚类。

④ 在**【输出】**栏中选择**【统计量】**和**【图】**，从而激活**【统计量】**和**【绘制】**两个按钮。

⑤ 选择要求输出的统计量。单击**【统计量】**按钮，打开**【系统聚类分析：统计量】**对话框，进行下列选择：

- **【合并进程表】**。要求输出凝聚状态表。
- **【相似性矩阵】**。要求输出距离矩阵。
- 在**【聚类成员】**栏中选择**【单一方案】**，并在**【聚类数】**框中输入“4”，即要求聚类进行到把所有观测分为 4 类时，显示每个观测所属的类。
- ⑥ 选择聚类方法。单击主对话框中的**【方法】**按钮，打开**【系统聚类分析：方法】**对话框。
  - 在**【聚类方法】**下拉列表中选择**【最远邻元素】**法。
  - 在**【度量标准】**栏中选择**【区间】**，因为聚类变量均为连续变量，即等间隔测度的变量。在下拉列表中选择**【平方 Euclidean 距离】**，即对等间隔测度的变量使用欧氏距离平方作为类间距离。
  - 在**【转换值】**栏中选择**【标准化】**方法。在标准化列表中选择**【全距从 0 到 1】**；选择**【按照变量】**项，即对每个等间隔测度的变量转换为全部值为 0~1 的范围。

注意：必须指定标准化方法，因为 4 个分析变量单位不同。

⑦ 选择要求显示的统计图。在主对话框的**【输出】**栏中选择**【图】**，单击**【绘制】**按钮，打开**【系统聚类分析：图】**对话框，进行下列选择：

- 选中**【树状图】**。
- 要求作冰柱图，但不要求把聚类全过程都表现在图上，而是只表现聚为 4 类的过程。因此在**【冰柱】**栏中选择**【聚类的指定全距】**项，并分别输入数字：**【开始聚类】**框中输入“1”，**【停止聚类】**框中输入“4”，**【排序标准】**框中输入“1”。在**【方向】**栏内选择**【垂直】**项，即纵向作图。

⑧ 选择要存入数据文件的新变量。在主对话框中单击【保存】按钮，在打开的对话框中选择【聚类成员】框中的【单一方案】项，在【聚类数】框中输入“4”，即要求在工作数据文件中建立新变量，当把所有观测分为 4 类时，该变量值表明每个观测被分派到的类号。

在主对话框中单击【确定】按钮，直接提交运行。

(2) 在输出窗中的输出结果见表 13-14~表 13-16 和图 13-24、图 13-25。

(3) 输出结果解释。

表 13-14 所示是欧氏不相似性系数平方矩阵，它是 20×20 方阵。第一行、第一列均是啤酒名，行列交叉点上这是两种啤酒 4 个变量的欧氏距离的平方和，体现的是不相似性，数值越大，两种啤酒越不相似。如果读者使用该数据集数据进行同样的分析，表中内容会稍有不同，这是因为本书作者对视图窗口中的表格进行了编辑，列宽的调整有时会影响显示的有效位数。

表 13-14 欧氏不相似性系数平方矩阵

案例	平方 Euclidean 距离																			
	1: Budweiser	2: Schlitz	3: Ionenbrau	4: Kronenbourg	5: Heineken	6: Old-milwaukee	7: Auschberger	8: Strichs-bohemi	9: Miller-lite	10: Sudeisler-lich	11: Coors	12: Coorslicht	13: Michelos-lich	14: Secrs	15: Kkirin	16: Pabst-extra-l	17: Hamm	18: Heilemans-old	19: Olympia-gold-	20: Schlite-light
1: Budweiser	.000	.111	.062	.724	.570	.140	.198	.147	.358	.556	.023	.213	.193	.391	.855	1.069	.014	.061	1.109	.530
2: Schlitz	.111	.000	.090	.665	.623	.249	.098	.230	.745	.886	.161	.591	.376	.467	.926	1.714	.183	.164	1.708	.933
3: Ionenbrau	.062	.090	.000	.390	.339	.337	.267	.348	.364	.482	.039	.301	.123	.323	.532	1.332	.104	.206	1.142	.475
4: Kronenbourg	.724	.665	.390	.000	.071	1.451	1.05	1.308	.815	.776	.589	.885	.418	.385	.054	2.269	.800	1.037	1.531	.756
5: Heineken	.570	.623	.339	.071	.000	1.272	.936	1.026	.682	.729	.471	.653	.345	.155	.059	1.899	.612	.801	1.331	.656
6: Old-milwaukee	.140	.249	.337	1.451	1.272	.000	.222	.130	.661	.930	.228	.457	.555	.929	1.672	1.162	.149	.114	1.497	.934
7: Auschberger	.198	.098	.267	1.054	.936	.222	.000	.137	1.041	1.358	.326	.805	.709	.630	1.354	2.086	.297	.114	2.239	1.314
8: Strichs-bohemi	.147	.230	.348	1.308	1.026	.130	.137	.000	.867	1.201	.283	.540	.643	.557	1.496	1.416	.168	.027	1.786	1.152
9: Miller-lite	.358	.745	.364	.815	.682	.661	1.04	.867	.000	.087	.222	.065	.122	.791	.741	.540	.292	.638	.288	.027
10: Sudeisler-lich	.556	.886	.482	.776	.729	.930	1.36	1.201	.087	.000	.363	.210	.132	.953	.703	.556	.473	.951	.196	.050
11: Coors	.023	.161	.039	.589	.471	.228	.326	.283	.222	.363	.000	.141	.087	.394	.685	.948	.026	.156	.873	.347
12: Coorslicht	.213	.591	.301	.885	.653	.457	.805	.540	.065	.210	.141	.000	.128	.572	.823	.443	.139	.388	.395	.148
13: Michelos-lich	.193	.376	.123	.418	.345	.555	.709	.643	.122	.132	.087	.128	.000	.428	.434	.810	.167	.455	.538	.153
14: Secrs	.391	.467	.323	.385	.155	.929	.630	.557	.791	.953	.394	.572	.428	.000	.395	1.695	.412	.451	1.496	.870
15: Kkirin	.855	.926	.532	.054	.059	1.672	1.35	1.496	.741	.703	.685	.823	.434	.395	.000	2.088	.893	1.199	1.283	.641
16: Pabst-extra-l	1.07	1.714	1.33	2.269	1.899	1.162	2.09	1.416	.540	.556	.948	.443	.810	1.69	2.07	.000	.847	1.314	.256	.607
17: Hamm	.014	.183	.104	.800	.612	.149	.297	.168	.292	.473	.026	.139	.167	.412	.893	.847	.000	.086	.927	.455
18: Heilemans-old	.061	.164	.206	1.037	.801	.114	.114	.027	.638	.951	.156	.388	.455	.451	1.20	1.314	.086	.000	1.535	.882
19: Olympia-gold-	1.11	1.708	1.14	1.531	1.331	1.497	2.24	1.786	.288	.196	.873	.395	.538	1.50	1.28	.256	.927	1.535	.000	.217
20: Schlite-light	.530	.933	.475	.756	.656	.934	1.31	1.152	.027	.050	.347	.148	.153	.870	.641	.607	.455	.882	.217	.000

这是一个对称相似矩阵

表 13-15 所示是系统聚类过程的输出。由于在统计量选项中选择了合并进程表，输出在输出窗中为一个表明聚类过程的表，其中：

- 阶是聚类步序号。群集组合中的群集 1、群集 2 是该步被合并的两类中的观测号。
- 系数。距离测度值。表明不相似性的系数。由于选择了欧氏距离平方作为距离测度，因此从表中可以看出，数值较小的两项(两个观测、两类或观测与一类)比数值较大的两项先合并。第一步是第 1 个观测与第 17 个观测合并；第二步是第 1 个和第 11 个观测合并。这样两步合并了 3 个观测到一类。
- 首次出现阶集群式指合并的两项第一次出现的聚类步序号。集群 1 和集群 2 值均为“0”的是两个观测合并，其中有一个为 0 的是观测与类合并；两个值均为非 0 值的是两个类合并。例如，第 6 步为第 4 观测与第 5 观测合并，而第 5 观测在第 5 步已经与第 15 观测合并为一类了。因此，此项值的 5 表示观测 4 与第 5 步形成的类归并为一类。
- 下一阶。此步合并结果继续在下一步合并时的步序号。
- 选择不同的标准化算法，距离矩阵会有所不同。选择不同的算法和对距离的测度方法，聚类过程是不同的，因而聚类结果也会有所区别。



图 13-25 所示是聚类全过程的树形图。可以在此图上用一把尺子垂直放在图上左右移动，与尺子相交的每根横线就是一类，每根横线左端与之联系的各观测就是分到该类的成员。大致观察一下，决定如何分类合适。图上方的数字是按距离比例进行重新标定的结果，不影响对分类结果的观察与结论。可以看出，分为 2 类、3 类或 4 类时，类间距离比较大，说明各类的特点比较突出，对各类啤酒容易定义；分为 5 类以上时，有些类间的区别不很明显。

图 13-26 所示是工作数据文件的一部分，其中最右列是新变量 `clu4_1`，其值表明聚为 4 类时各观测所属类的类号。

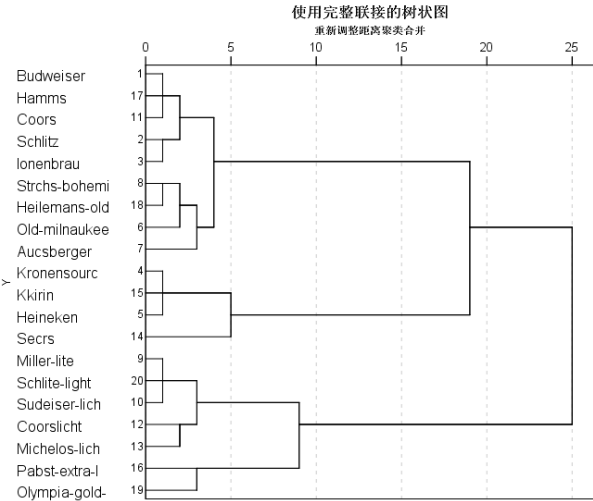


图 13-25 聚类树形图(重新标定到 0~25)

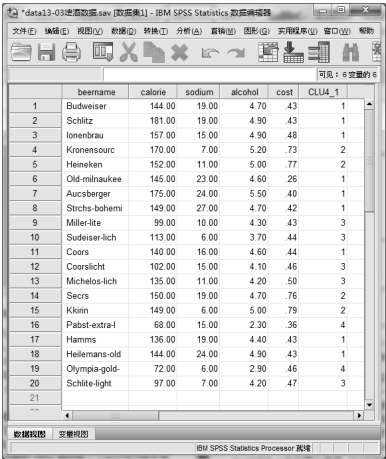


图 13-26 加入了新变量的工作数据文件

比较表 13-16 与图 13-26 可以看出，在工作数据文件中建立的新变量也表明各观测所属的类，因此可以根据需要选择一个即可。若在【系统聚类分析：保存】对话框中指定了建立新变量，可以不必在【系统聚类分析：统计量】对话框中指定【聚类成员】的选项。

比较时应该注意，变量 `clun_1` 的值只是序号。采用相同的  $n$  值，不同的聚类方法和不同的测度不相似性(距离)的算法，结果可能有区别。例如，对于都是聚为 4 类，第 1 次、第 2 次、第 3 次执行采用不同聚类方法和不同测度距离的方法，会建立 `clu4_1`、`clu4_2`、`clu4_3`、... 的变量，可以比较结果，得出结论。

【例 5】 【例 4】使用另一些选项的程序与输出。

应该说明的是，分类是根据特定的目的进行的。对于同样一些观测，不同的分类目的，使用反映不同特征的变量，分类的结果就不相同。同一分类目的，根据不同的实际需要，也可以分成不同的类数。因此，可以在使用系统聚类时指定不同的参数，对不同的结果进行比较，以便得出符合实际需要的结论。

1) 操作步骤

(1)~(4) 步操作与【例 4】相同。

(5) 选择显示的统计量。在【系统聚类分析：统计量】对话框中【合并进程表】项和【相似性矩阵】项的选择与【例 4】相同。

在【聚类成员】栏中选择【方案范围】项，并在【最小聚类数】框中输入“3”、【最大聚类数】框中输入 5，即要求聚类进行到把所有观测分为 3 类、4 类、5 类时，显示每个观测所属的类。



(6) 在【系统聚类分析：方法】对话框中，选择【聚类方法】为【质心聚类】法；在【标准化】框中，选择【Z 得分】。

(7) 选择要求显示的统计图。选择【树状图】，要求作树形图。选择【冰柱图】，要求作冰柱图，但不要求把聚类全过程都表现在图上，而是只表现聚为 4 类的过程，因此在【冰柱】栏中选择【聚类的指定全距】项，并分别输入数字：【开始聚类】框中输入“3”，【停止聚类】框中输入“10”，【排序标准】框中输入“2”，即要求作冰柱图表明聚为 3 类、5 类、7 类、9 类时的分类情况。在【方向】栏中仍选择【垂直】项，即作纵向冰柱图。

(8) 选择要存入数据文件的新变量。  
在主对话框中单击【保存】按钮，打开相应的对话框。选择【聚类成员】框中的【方案范围】项，在【最小聚类数】框中输入“2”，【最大聚类数】框中输入“6”，即要求在工作数据文件中建立 5 个新变量。当把所有观测分为 2 类、3 类、4 类、5 类和 6 类时对应变量值表明每个观测被分派到的类号。

2) 结果输出 (见表 13-17、图 13-27 和图 13-28)

3) 结果解释

(1) 选择的计算不相似性系数(距离)的方法与【例 4】相同，但标准化方法不同，因此不相似性矩阵输出结果与表 13-14 也不同。聚类方法选择与【例 4】相同，但由于标准化方法不同聚类的凝聚过程与表 13-15 也不同。为节省篇幅，请读者自行观察这两项输出。

聚类结果见表 13-17。左表与右表分别是使用不同标准化的方法产生的不同结果，都是聚为 3 类、4 类、5 类的结果，但聚类结果是有差别的。左表是用 Z 分数方法标准化的聚类结果；右表是用标准化到 0~1 的方法的聚类结果。读者可以自己继续比较。

表 13-17 不同标准化方法聚为 3 类、4 类、5 类的结果

群集成员				群集成员			
案例	5 群集	4 群集	3 群集	案例	5 群集	4 群集	3 群集
1: Budweiser	1	1	1	1: Budweiser	1	1	1
2: Schlitz	1	1	1	2: Schlitz	1	1	1
3: Ionenbrau	1	1	1	3: Ionenbrau	1	1	1
4: Kronensourc	2	2	2	4: Kronensourc	2	2	2
5: Heineken	2	2	2	5: Heineken	2	2	2
6: Old-milnaukee	1	1	1	6: Old-milnaukee	1	1	1
7: Aucsberger	1	1	1	7: Aucsberger	1	1	1
8: Strchs-bohemi	1	1	1	8: Strchs-bohemi	1	1	1
9: Miller-lite	3	3	1	9: Miller-lite	3	3	3
10: Sudeiser-lich	3	3	1	10: Sudeiser-lich	3	3	3
11: Coors	1	1	1	11: Coors	1	1	1
12: Coorslicht	3	3	1	12: Coorslicht	3	3	3
13: Michelos-lich	3	3	1	13: Michelos-lich	3	3	3
14: Secrs	4	2	2	14: Secrs	4	2	2
15: Kkirin	2	2	2	15: Kkirin	2	2	2
16: Pabst-extra-l	5	4	3	16: Pabst-extra-l	5	4	3
17: Hamms	1	1	1	17: Hamms	1	1	1
18: Heilemans-old	1	1	1	18: Heilemans-old	1	1	1
19: Olympia-gold-	5	4	3	19: Olympia-gold-	5	4	3
20: Schlite-light	3	3	1	20: Schlite-light	3	3	3

(2) 图 13-27 所示为使用 Z 分数法对原始变量进行标准化，反映聚类过程的冰柱图。从冰柱图可以很清晰地看出如果分 3 类，用观测序号则表示：

第 1 类包括的是编号为 1、2、3、6、7、8、11、17、18 的啤酒。

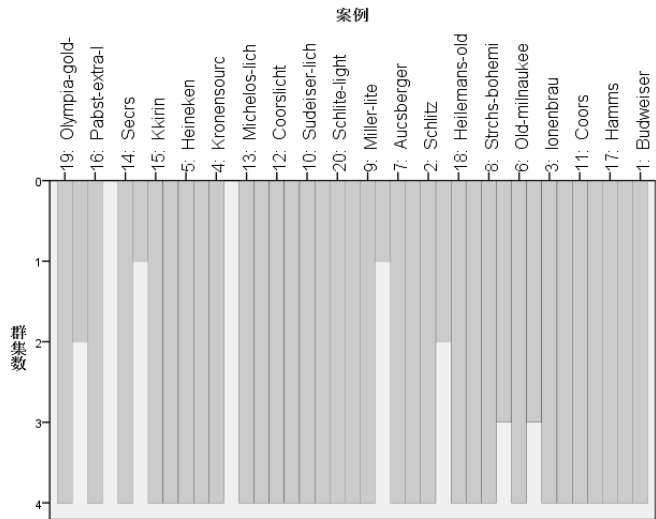


图 13-27 第 3、5、7、9 步聚类的纵向冰柱图

第 2 类包括的是编号为 4、5、9、10、12、13、14、15、20 的啤酒。  
第 3 类包括的是编号为 19、16 的啤酒。  
如果分为 5 类，第 2 类分成 2 类，则编号为 4、5、15、14 和 9、20、10、12、13 的啤酒各聚成单独的一类。第 1 类分为 1、17、11、2、3 和 6、8、18、7 两类。  
以此类推，读者可自行观察分为 7 类或分为 9 类的各啤酒的分类结果。对照不相似性系数矩阵，会对聚类的原理有更进一步的理解。  
(3) 图 13-28 所示是在工作数据文件中建立的新变量，共 5 个。

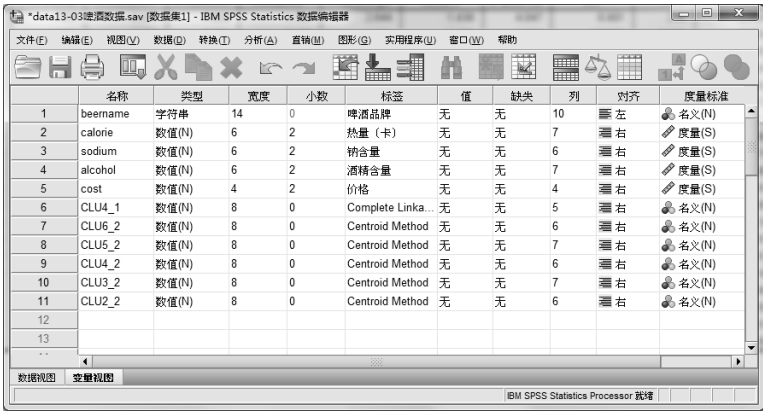


图 13-28 数据窗中生成的类别新变量（变量视图）

图 13-28 所示是变量视图，可以看到系统自动命名的新变量。此次运行建立的新变量的变量名 clu6\_2、clu5\_2、clu4\_2、clu3\_2、clu2\_2 分别表示当前的 SPSS 期间第二次运行系统聚类过程，如果分为 6 类或 5 类、4 类、3 类、2 类时各观测所属类别变量。  
图 13-29 所示是数据文件窗口的数据视图。每个变量的值为相应的观测分到类的代码。这里的第几类没有任何含义，只是标记。至于哪类是什么特征，还需认真分析工作数据窗口中的原始数据、新生成的分类变量，根据专业知识来确定，并可以进一步对每一类进行命名。

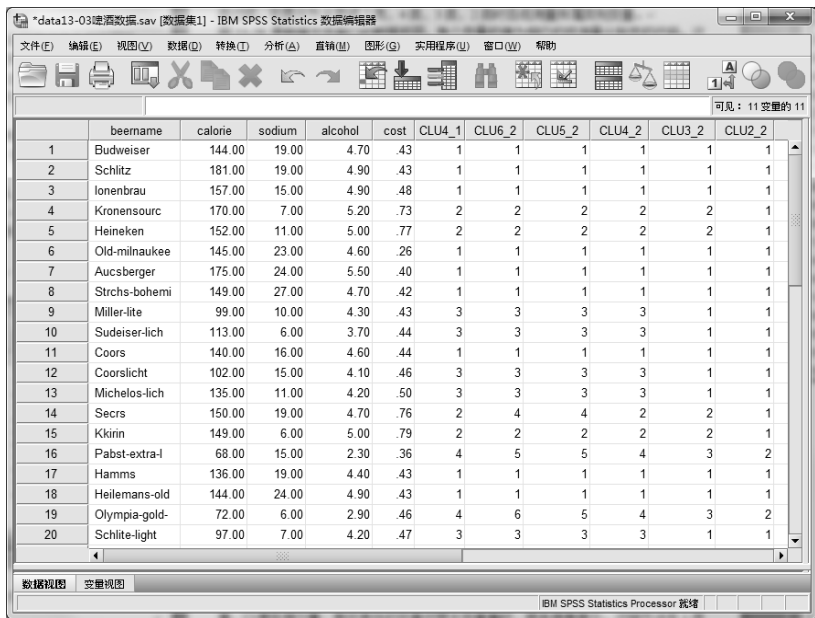


图 13-29 数据窗中生成的类别新变量(数据视图)

注意：SPSS 提供了众多的聚类方法和标准化方法。分析数据时，都是人为选定某种方法。不同的聚类算法和不同的对变量进行标准化的方法都会对聚类结果有影响。而类本身就是针对某一研究目的而进行的。因此在作结论时，一定要结合专业知识、研究目的，同时认真观察原始数据特征，审慎地得出结论，并对分成的各类进行命名。如果不同方法得出的结果差别很大，说明聚类变量选择的不是真正反映观测的分类特征，应该在变量选择上下工夫。

图 13-30 所示是质心联接法聚类的树形图。读者可以对照图 13-29 的数据，自己分析聚成几类比较好。

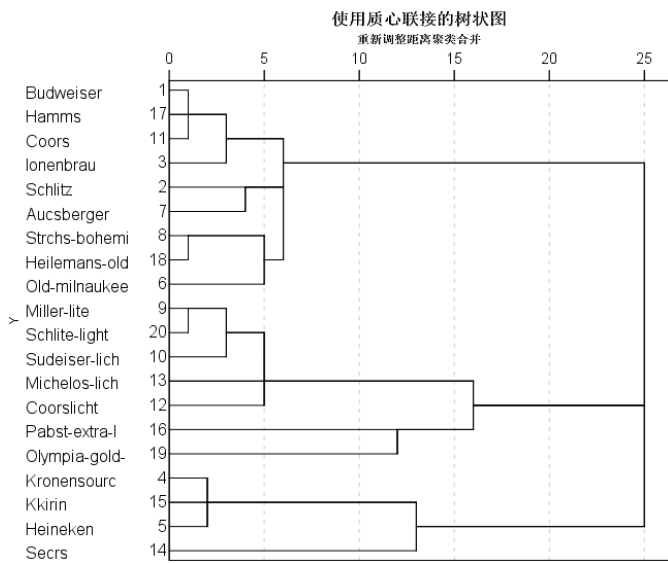


图 13-30 使用质心联接法聚类的树形图

## 13.4.4 变量聚类概述

### 1. 变量聚类的概念

变量聚类也称 R 型聚类,是一种降维的方法,用于在变量众多时寻找有代表性的变量,以便在用少量、有代表性的变量代替大变量集时,损失信息很少。这种方法在人类学、动物学、医学和工业生产中以及市场分析中都得到应用,例如人种分类、动植物分类等,往往要测量许多表明形态特性的变量值。某些变量之间有很强的相关性,找出一个变量可以代替一系列与其相关的变量,则可大大减少工作量,节省测量时间,但不会影响分类的结果。因此,在分类学中选择变量是一步很重要的工作。变量聚类是选择变量的很实用的方法之一。另外,进行回归分析时也需要首先降维以便找出相互独立的变量。

### 2. 选择代表指标的方法

聚类结束后,各类变量中选择哪个变量作为代表变量呢?典型指标的选择主要根据专业知识,同时根据下列原则综合确定代表变量,考察在一类指标中:

(1) 最有代表性的变量。

(2) 最容易测得的变量。例如,测试仪器容易得到、仪器便宜、测试对象容易接受、指标数据容易测得准确等各方面因素。例如医学研究中,尿量虽然容易测得,但 24h 尿量不易收全,就不易准确。此时就应该考虑,在与之聚为一类的变量中,其他变量中是否有更好的代替者。

(3) 如果从专业角度不好确定,还可以通过进行进一步计算来确定。

例如,  $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$  这 4 个指标已经根据 R 型聚类结果聚为一类。

① 计算每个指标的相关指数,公式为

$$\bar{R}_j^2 = \frac{\sum r^2}{m_j - 1}$$

式中,  $r$  为指标  $x_j$  与同类中其他指标间的相关系数;  $m_j$  为指标  $x_j$  所在类的指标个数。

② 对  $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$  这 4 个指标计算  $\bar{R}_1^2$ 、 $\bar{R}_2^2$ 、 $\bar{R}_3^2$ 、 $\bar{R}_4^2$ ,比较这 4 个值,最大一个相关指数对应的变量,可以选作典型指标。

### 3. SPSS 使用聚类过程对变量进行聚类

操作步骤与方法均与使用聚类过程对观测进行聚类是相同的。不同点在于:

(1) 在【系统聚类分析】主对话框中的【分类】框中选择【变量】项。

(2) 【系统聚类分析】主对话框中的【保存】按钮为灰色,不能单击。因为变量聚类不建立新变量。

## 13.4.5 变量聚类分析实例

**【例 6】**啤酒分类的问题中,是否有必要使用 4 个变量进行分析?可以用变量聚类方法解决这个问题。数据文件仍为 data13-03。

1) 操作步骤

(1) 按【分析→分类→系统聚类】顺序单击菜单项,打开主对话框。

(2) 选择 4 个变量:calorie(热量)、sodium(钠含量)、alcohol(酒精含量)、cost(价格)为分析变量,移到【变量】栏中。在【聚类】栏中选择【变量】项。

- (3) 单击【方法】按钮打开相应对话框。
- ① 在【聚类方法】下拉列表中选择【最远邻元素】作为聚类方法。
- ② 在【度量标准】栏中选择【区间】中的【Pearson 相关性】作为测度变量间相似性的方法。也因此在此【转换值】栏选择【标准化】中的【无】项。不进行标准化。
- (4) 单击【绘制】按钮，打开相应的对话框，选择【树状图】项。在【冰柱】栏中选择所有类，要求在冰柱图中反映聚类全过程。
- (5) 单击【统计量】按钮，打开相应的对话框。选择【相似性矩阵】。要求显示相关系数矩阵。在主对话框中单击【确定】按钮，提交系统执行。
- 2) 运行结果(见表 13-18、表 13-19 和图 13-31、图 13-32)
- 省略综合信息表和聚类过程表。
- 3) 输出结果解释

表 13-18 变量的相关矩阵  
近似矩阵

案例	矩阵文件输入			
	热量 (卡)	钠含量	酒精含量	价格
热量 (卡)	1.000	.429	.903	.291
钠含量	.429	1.000	.337	-.444
酒精含量	.903	.337	1.000	.345
价格	.291	-.444	.345	1.000

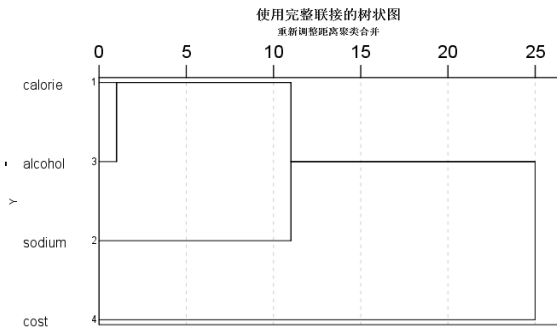


图 13-31 变量聚类的树状图

表 13-19 聚类过程表  
聚类表

阶	群集组合		系数	首次出现阶群集		下一阶
	群集 1	群集 2		群集 1	群集 2	
1	1	3	.903	0	0	2
2	1	2	.337	1	0	3
3	1	4	-.444	2	0	0

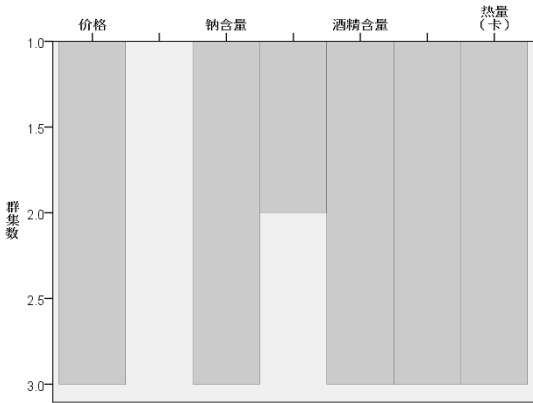


图 13-32 变量聚类的冰柱图

无论从相关矩阵还是冰柱图、树形图，都可以看出热量和酒精含量两个变量相关系数最大，首先聚为一类。从整体看，聚为 3 类是比较好的结果。至于“热量”和“酒精”含量选择哪一个作为典型指标代替原来的两个变量，可以根据专业知识或测定的难易程度决定。

**【例 7】** 为更好地说明选择典型变量的计算方法，再举一例。

有 10 个测验项目，分别用变量  $x_1 \sim x_{10}$  表示。50 名学生参加测试。数据文件为 data13-04。要求：对 10 个变量进行变量聚类；计算并打印各变量间的相关矩阵，用相关测度各变量间的距离。打印出聚为 2 类的结果，即各变量属于 2 类中的哪一类；打印出聚类全过程的冰柱图，以便对于变量分类进行进一步的探讨。

根据要求的操作步骤如下：

- (1) 读取数据文件 data13-04。按【分析→分类→系统聚类】顺序单击菜单项，打开【系统聚类分析】主对话框。

- (2) 在主对话框中指定分析变量，在变量表中选择  $x_1 \sim x_{10}$ ，移到【变量】框中。
- (3) 在【聚类】栏中选择【变量】项，即选择进行变量聚类。
- (4) 在主对话框中单击【方法】按钮，在打开的对话框中选择聚类方法：在【聚类方法】下拉列表中选择【最远邻元素】项，在【度量标准】栏内选择【区间】，在下拉列表中选择【Pearson 相关性】项，即皮尔逊相关作为测度变量间相关性的方法。
- (5) 在主对话框中单击【统计量】按钮打开相应的对话框，选择输出项：选择【相似性矩阵】项，要求打印相关矩阵。在【聚类成员】栏中，选择【单一方案】，并【聚类数】框中输入“2”。
- (6) 选择输出的统计图。在主对话框中，单击【绘制】按钮，打开相应的对话框。在【冰柱】栏中选择所有聚类项，即要求显示聚类全过程的冰柱图。
- (7) 在主对话框单击【确定】按钮，提交运行。
- (8) 在输出窗口中输出结果，见表 13-21、表 13-22 和图 13-33、图 13-34，其中略去了数据的综合信息。

表 13-20 变量聚类的相关系数矩阵

案例	矩阵文件输入									
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1.000	.133	.290	.099	.331	.198	.449	.323	.320	.112
x2	.133	1.000	.026	.411	.201	.328	.134	.199	.268	.271
x3	.290	.026	1.000	.151	.274	.406	.443	.509	.598	.318
x4	.099	.411	.151	1.000	.072	.282	.145	.401	.324	.407
x5	.331	.201	.274	.072	1.000	.317	.191	.063	.356	.084
x6	.198	.328	.406	.282	.317	1.000	.370	.312	.306	.296
x7	.449	.134	.443	.145	.191	.370	1.000	.337	.313	.246
x8	.323	.199	.509	.401	.063	.312	.337	1.000	.611	.584
x9	.320	.268	.598	.324	.356	.306	.313	.611	1.000	.325
x10	.112	.271	.318	.407	.084	.296	.246	.584	.325	1.000

表 13-21 聚类进程表

阶	群集组合		系数	首次出现阶群集		下一阶
	群集 1	群集 2		群集 1	群集 2	
1	8	9	.611	0	0	2
2	3	8	.509	0	1	5
3	1	7	.449	0	0	7
4	2	4	.411	0	0	8
5	3	10	.318	2	0	9
6	5	6	.317	0	0	7
7	1	5	.191	3	6	8
8	1	2	.072	7	4	9
9	1	3	.026	8	5	0

表 13-22 两类的类成员表

案例	2 群集
x1	1
x2	1
x3	2
x4	1
x5	1
x6	1
x7	1
x8	2
x9	2
x10	2

- (9) 输出结果说明。
- 表 13-20 所示是测度变量间距离的相关矩阵。
- 表 13-21 所示是聚类过程。可以看出自上至下相关系数是下降的。也就是说，相关系数大的先聚成一类。对比表 13-20，可以理解变量聚类的过程。
- 表 13-22 所示是聚为 2 类的结果，一类由  $x_1$ 、 $x_2$ 、 $x_4$ 、 $x_5$ 、 $x_6$ 、 $x_7$  组成，另一类由  $x_3$ 、 $x_8$ 、 $x_9$ 、 $x_{10}$  组成。

图 13-33 所示是聚类全过程的冰柱图，可以看出分成两类的结果与表 13-22 是一致的，还可以查看如果聚为 3 类，各类组成为  $x_{10}$ 、 $x_9$ 、 $x_8$ 、 $x_3$ ； $x_4$ 、 $x_2$ ； $x_6$ 、 $x_5$ 、 $x_7$ 、 $x_1$ 。若聚为 4 类，各类组成为  $x_{10}$ 、 $x_9$ 、 $x_8$ 、 $x_3$ ； $x_2$ 、 $x_4$ ； $x_5$ 、 $x_6$ ； $x_7$ 、 $x_1$ 。

图 13-34 所示是聚类的树形图。图形的横坐标重新标定到距离到 25。重新标定不影响对聚类结果的判断。如果以横坐标值为 20 划分，可以分为 3 类；如果按横坐标值 15 划分，就分成 4 类；如果以横坐标值为 10 划分就分为 6 类。在实际工作中完全是人为根据研究的需要，根据冰柱图和专业知识确定聚为几类最为合理，最后得出结论。

并非是客观存在的分类。

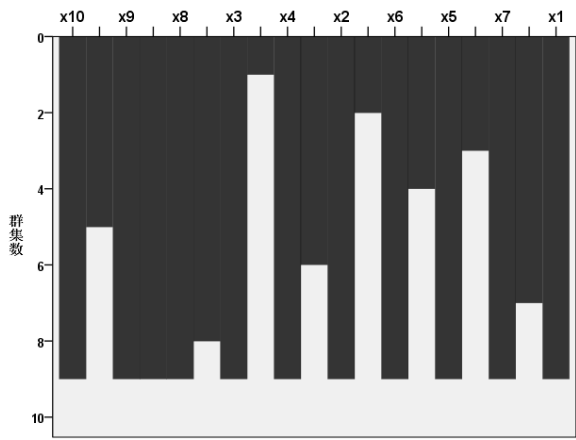


图 13-33 变量聚类全过程的冰柱图

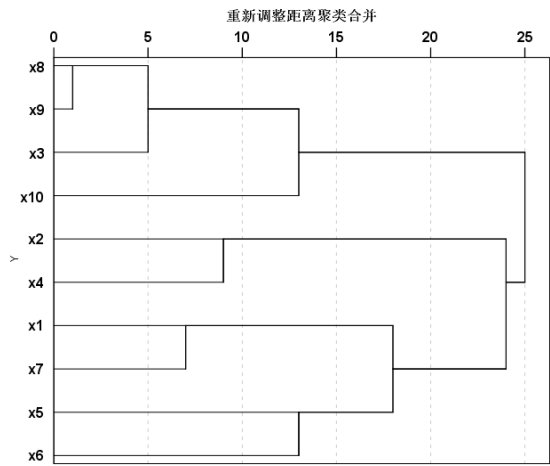


图 13-34 变量聚类的树状图(重新标定距离到 25)

(10) 典型指标的选择。

聚类结果说明只要有两个有代表性的变量就可以了，但是聚类结果每一类中都有两个以上的变量，选择哪个变量代表这一组变量呢？即每组都要选择具有代表性的典型变量。

根据下面的公式计算相关指数：

$$\bar{R}_j^2 = \frac{\sum r^2}{m_j - 1}$$

选择典型指标或称代表变量。以第一组为例，计算  $\bar{R}_{10}^2$ 、 $\bar{R}_9^2$ 、 $\bar{R}_8^2$ 、 $\bar{R}_3^2$ ，方法如下：

- ① 按【分析→相关→双变量】菜单顺序打开【双变量相关分析】对话框。
- ② 选择分析变量  $x_3$ 、 $x_8$ 、 $x_9$ 、 $x_{10}$ 。
- ③ 选择分析方法，在【相关系数】栏选择【Pearson】。

表 13-23 第一组变量相关矩阵

		x3	x8	x9	x10
x3	Pearson 相关性	1	.509**	.598**	.318
	显著性（双侧）		.000	.000	.025
	N	50	50	50	50
x8	Pearson 相关性	.509**	1	.611**	.584**
	显著性（双侧）	.000		.000	.000
	N	50	50	50	50
x9	Pearson 相关性	.598**	.611**	1	.325
	显著性（双侧）	.000	.000		.021
	N	50	50	50	50
x10	Pearson 相关性	.318	.584**	.325	1
	显著性（双侧）	.025	.000	.021	
	N	50	50	50	50

\*\* . 在 .01 水平（双侧）上显著相关。  
\* . 在 0.05 水平（双侧）上显著相关。

- ④ 单击【确定】按钮提交运行，得到表 13-23 所示的相关矩阵表。

从表中读取相关系数，计算各相关指数：

$$\bar{R}_3^2 = (0.509^2 + 0.598^2 + 0.318^2) / 3 = 0.23927$$
$$\bar{R}_8^2 = (0.509^2 + 0.611^2 + 0.584^2) / 3 = 0.32449$$
$$\bar{R}_9^2 = (0.598^2 + 0.611^2 + 0.325^2) / 3 = 0.27885$$
$$\bar{R}_{10}^2 = (0.318^2 + 0.584^2 + 0.325^2) / 3 = 0.18260$$

比较 4 个相关指数， $x_8$  的相关指数最大，

因此该组变量选择  $x_8$  作代表变量。其余各组的代表变量读者可以自行按上述方法计算。

在科学研究中，除了根据计算出的相关指数，还要考虑哪个变量的值容易获取，哪个变量的精度容易保证等因素，综合考虑代表变量的选择。

## 13.5 判别分析

### 13.5.1 判别分析概述

#### 1. 判别分析的概念

判别分析是一种常用的统计分析方法。判别分析是根据观察或测量到若干变量值，判断研究对象属于哪一类的方法。例如，医学实践中根据各种化验结果、疾病症状、体征判断患者患的是什么疾病；体育人才选拔是根据运动员的体形、运动成绩、生理指标、心理素质指标、遗传因素判断是否选入运动队继续培养；动物、植物分类等都可以用判别分析来解决。判别分析是应用计算机进行运动员选才、动物、植物分类以及疾病辅助诊断等的主要统计学基础。

进行判别分析必须已知观测对象的分类和若干表明观测对象特征的变量值。判别分析就是要从中筛选出能提供较多信息的变量并建立判别函数，使得利用推导出的判别函数对观测判别其所属类别时的错判率最小。

线性判别函数的一般形式为

$$y = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n$$

式中， $y$  为判别分数（判别值）； $x_1, x_2, x_3, \cdots, x_n$  为反映研究对象特征的变量； $a_1, a_2, a_3, \cdots, a_n$  为各变量的系数，也称判别系数。

SPSS 对于分为  $m$  类的研究对象，建立  $m$  个线性判别函数。对每个个体进行判别时，把测试的各变量值代入判别函数，得出判别分数，或者计算属于各类的概率，从而确定该个体属于哪一类；还建立标准化和未标准化的典则判别函数。



## 2. 判别分析过程的功能

SPSS 提供的判别分析过程是根据已知的观测分类和表明观测特征的变量值推导出判别函数,并把各观测的自变量值回代到判别函数中,根据判别函数对观测所属类别进行判别。对比原始数据的分类和按判别函数所判的分类,给出错分概率。

判别分析可以根据类间协方差矩阵,也可以根据类内协方差矩阵进行分析。如果原始数据中的观测是分为  $m$  类的,每一已知类的先验概率可以取其值相等,即等于  $1/m$ ,也可以与各类样本量成正比。

判别分析可以根据要求,给出各类观测的单变量的描述统计量、线性(费雪 Fisher)判别函数的系数或标准化及未标准化的典则判别函数的系数、类内相关矩阵、类内和类间协方差矩阵和总协方差矩阵,给出按判别函数判别(回代)的各观测所属类别,带有错分率的判别分析小结,还可以根据要求生成表明各类分布的区域图和散点图。如果希望把部分聚类结果存入文件,还可以在工作数据文件中建立新变量,表明观测按判别函数分派的类别、按判别函数计算的判别分数和分到各类去的概率。

判别分析过程的大部分功能都可以通过对话框来指定,还有一些功能可以在语句窗口中给予补充或修改。例如指定各类的先验概率、显示旋转方式和结构矩阵、限制提取的判别函数的数目、读取一个相关矩阵、分析后把相关矩阵写入文件、指定对参与分析的观测进行回代分类,对没有参与分析的观测进行预测分类等。可参见第 13.5.4 节相关内容。

## 3. 有关判别分析的术语

(1) 建立判别函数的方法有 4 种:全模型法、向前选择法、向后选择法、逐步选择法。

① 全模型法。全模型法是把读者指定的变量全部放入判别函数中,不管变量对判别函数是否起作用、作用大小如何。当对反映研究对象特征的变量认识比较全面时可以选择此种方法。此种方法是 SPSS 系统默认的方法。

由于人们对客观事物的认识可能并不客观,因此对变量的选择就有可能出现偏差。如果没有选择对研究对象的特征能够提供丰富信息的变量,没有测试有关的数据,只能等待人们对所研究事物认识的进一步深化,别无他法。但是,如果选择的变量中有对研究对象的特征不能提供较丰富的信息,对判别贡献很小的变量,这样的变量就应该从判别模型中剔除。全模型方法不能解决这个问题。

② 向前选择法。此方法是从判别模型中没有变量开始,每一步把一个对判别模型的判断能力贡献最大的变量引入模型,直到在没有被引入模型的变量中没有一个符合进入模型的条件(判据)时,变量引入过程结束。当希望比较多的变量留在判别函数中时,使用向前选择法。

③ 向后选择法。此方法与向前选择法完全相反,它是把用户所有指定的变量建立一个全模型,每一步把一个对模型的判断能力贡献最小的变量剔除出模型,直到模型中的所有变量都符合留在模型中的判据时,剔除变量工作结束。在希望较少的变量留在判别函数中时使用向后选择法。

④ 逐步选择法。此判别法从模型中没有变量开始,每一步都要对模型进行检验。每一步都在把模型外的对模型的判断能力贡献最大的变量加入到模型中的同时,也考虑把已经在模型中但又不符合留在模型中的条件的变量剔除。这是因为新变量的引入有可能使原来已经在模型中的变量对模型的贡献变得不显著了。直到模型中的所有变量都符合引入模型的判据,模型外的变量都不符合进入模型的判据时,逐步选择变量的过程停止。逐步选择法更能比较好地选择变量,SPSS 用此种方法建立非全(变量)判别函数。此种方法作为可选择的方法。

(2) 典则判别分析。典则判别分析建立典则变量代替原始数据文件中指定的自变量。典则变量是原始自变量的线性组合。用少量的典则变量代替原始的多个变量可以比较方便地描述各类之间的关系，如可以用平面区域图或散点图直观地表示各类之间的相对关系。SPSS 计算标准化和未标准化的典则判别函数系数。

(3) 判别函数的性能。判别分析得出的判别函数性能如何，可以通过回代的方法进行验证。即将各观测的变量值代到线性判别函数中，根据线性判别函数值(判别分数)确定每个观测分属于哪一类，然后与原始数据中的分类变量值进行比较，得到错判率。错判率越小说明判别函数的判别性能越好。

(4) 判别分析对数据的要求。进行判别分析要求数据遵循多元正态分布。实践工作中收集的数据，其分布往往不同于正态分布，因此使用本节介绍的参数分析方法是不合适的。从非正态总体导出的线性判别函数(或经过预处理的数据)导出的二次判别函数的误差率估计可能会有较大的偏差。

(5) 利用判别函数对观测进行分类。用判别分析过程导出的线性判别函数的数目与类别数目相同。确定一个观测属于哪一类，可以把该观测的各变量值代入每个判别函数，哪个判别函数数值大，该观测就属于哪一类。

13.5.2 判别分析过程

1. 建立或读入数据文件

在数据窗口中输入待分析的数据或利用文件菜单中的打开命令打开已经存在的数据文件，显示到数据窗口中。数据中必须包括一个表明已知的观测所属类别的变量和若干个表明分类特征的变量。

2. 打开主对话框

按【分析→分类→判别】顺序单击菜单项，打开【判别分析】主对话框，见图 13-35。

3. 选择分类变量及其范围

在主对话框的源变量表中选择表明已知的观测所属类别的变量(一定是离散型变量)送入【分组变量】框中。此时矩形框下面的【定义范围】按钮加亮，单击该按钮，打开【判别分析：定义范围】对话框，见图 13-36。在相应的框中输入该分类变量的最小值、最大值。

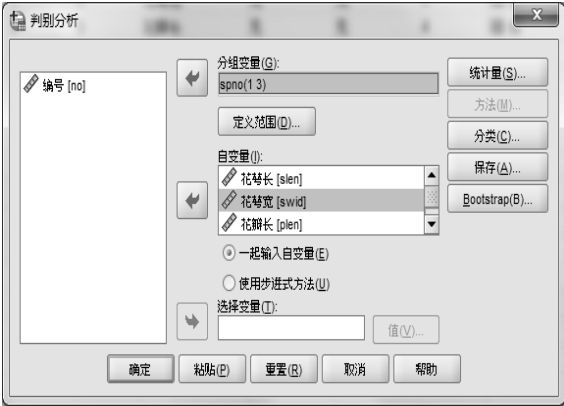


图 13-35 【判别分析】主对话框

4. 指定判别分析的自变量

在主对话框的源变量表中选择表明观测特征的变量，送到【自变量】框中，作为参与判别分析的变量。

5. 进行判别分析

完成前面 4 步的操作，即可使用系统默认值对工作数据集的数据进行判别分析了。单击【确定】按钮提交执行，在输出窗中显示出分析结果。

完全使用系统默认值进行判别分析，结果有时不能令人满意，因此根据以下步骤指定选项是很有必要的。

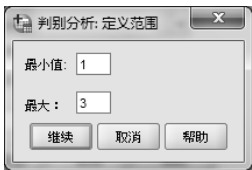


图 13-36 【判别分析: 定义范围】对话框

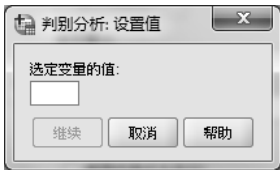


图 13-37 【判别分析: 设置值】对话框

6. 选择观测

如果希望使用一部分观测进行判别函数的推导, 而且有一个变量的某个值可以作为这些观测的标识, 则可以在主对话框中从源变量表框中选择这个变量送入【选择变量】栏中, 再单击【值】按钮, 打开子对话框, 见图 13-37, 输入标识参与分析的观测所具有的该变量值。使用数据中的所有合法观测, 此步骤可以省略。

7. 选择分析方法

在主对话框中的自变量矩形框下面有两个选项, 可从中选择判别分析方法。

(1) 【一起输入自变量】。当认为所有自变量都能对观测的特性提供丰富的信息, 且彼此独立时使用该选项。判别分析过程将不加选择地使用所有自变量进行判别分析, 建立全模型, 不需要进一步进行选择。

(2) 【使用步进式方法】。当不认为所有自变量都能对观测的特性提供丰富的信息时, 使用该选项。选择该项, 【方法】按钮加亮。

单击【方法】按钮, 打开【判别分析: 步进法】对话框, 见图 13-38。

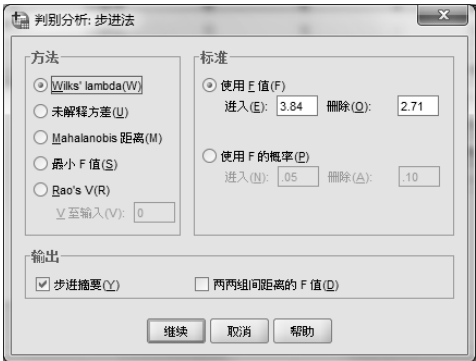


图 13-38 【判别分析: 步进法】对话框

- ① 在【方法】栏中可供选择的判别分析方法有:
- 【Wilks' lambda】。每步都是 Wilk 的  $\lambda$  统计量最小的进入判别函数。
  - 【不(未)解释方差】。每步都是各类不可解释的方差和最小的变量进入判别函数。
  - 【Mahalanobis 距离】。每步都使靠得最近的两类间的 Mahalanobis 距离最大的变量进入判别函数。
  - 【最小 F 值】。每步都是使任何两类间最小的  $F$  值最大的变量进入判别函数。
  - 【Rao's V】。每步都是使 Rao's V 统计量产生最大增量的变量进入判别函数。可以对一个要加入到模型中的变量的  $V$  值指定一个最小增量。选择此种方法后, 应该在该项下面的【V 至输入】框中输入这个增量的指定值。当某变量导致的  $V$  值增量大于指定值时, 该变量进入判别函数。

- ② 在【标准】栏中选择逐步判别停止的判据。可供选择的判据有:
- 【使用 F 值】。是系统默认的判据。当加入一个变量(或删除一个变量)后, 对在判别函数中的变量进行方差分析。当计算的  $F$  值大于指定的进入值时, 该变量保留在函数中, 默认值是进入值为“3.84”。当该变量使计算的  $F$  值小于指定的删除值时, 该变量从函数中剔除, 默认值是删除值为“2.71”。设置这两个值时应该注意使进入值大于删除值, 否则将致使生成的函数中没有变量。

- **【使用 F 的概率】**。即用检验的概率决定变量是否加入到函数中或被剔除。进入模型(或保留在模型内)的变量的  $F$  值概率的默认值是 0.05 (5%)，删除变量(不进入)的  $F$  值概率是 0.10 (10%)。应该保证删除变量的  $F$  值概率(或移出变量的  $F$  值概率)大于进入(加入、保留)变量的  $F$  值概率。
- ③ 在**【输出】**栏中的两项可以选择要显示的统计量。
- **【步进摘要】**。要求在逐步选择过程中的每一步之后显示每个变量的统计量。
- **【两两组间距离的 F 值】**。要求显示两两类之间的两两  $F$  值矩阵。

8. 指定输出的统计量

单击**【统计量】**按钮，打开**【判别分析：统计量】**子对话框，见图 13-39。

(1) 在**【描述性】**栏中选择要输出的原始数据的描述统计量。

- ① **【均值】**。输出各类中各自变量均值、标准差和各自变量总样本的均值和标准差。
- ② **【单变量 ANOVA】**。要求进行假设检验，输出单变量方差分析结果。检验的无效假设是：各类中同一自变量均值都相等。
- ③ **【Box's M】**。对各类的协方差矩阵相等的假设进行检验。如果样本足够大，则表明差异不显著的  $p$  值意味着矩阵差异不明显。

(2) 在**【函数系数】**栏中选择判别函数中各自变量系数(也称判别系数)的输出形式。

- **【Fisher】**。要求输出可以直接用于对新样本进行判别分类的费雪系数。对每一类给出一组系数，并给出该组中判别分数最大的观测。
- **【未标准化】**。要求输出未经标准化的判别系数。

(3) 在**【矩阵】**栏中选择要求给出的自变量系数矩阵。

- **【组内相关】**。要求输出类内相关矩阵，它是根据计算相关矩阵之前，将各组(类)协方差矩阵平均后计算的。
- **【组内协方差】**。要求计算并显示合并类内协方差矩阵，是将各组(类)协方差矩阵平均后计算的，区别于总协方差矩阵。
- **【分组协方差】**。对每类输出并显示一个协方差矩阵。
- **【总体协方差】**。计算并显示总样本的协方差矩阵。

9. 指定分类参数和判别结果

在主对话框中单击**【分类】**按钮，打开**【判别分析：分类】**对话框，见图 13-40。

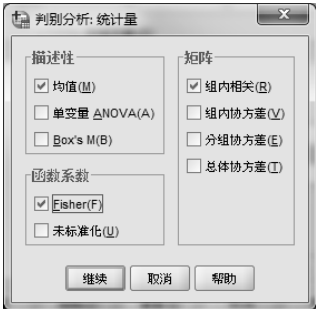


图 13-39 【判别分析：统计量】对话框

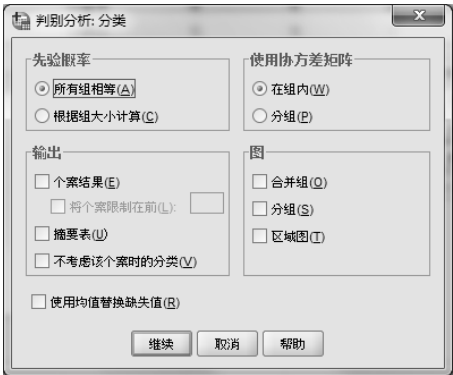


图 13-40 【判别分析：分类】对话框

(1) 在【先验概率】栏中选择先验概率。

- 【所有组相等】。各类先验概率相等。若分为  $m$  类，则各类先验概率均为  $1/m$ 。
- 【根据组大小计算】。各类的先验概率与各类的样本量成正比。

(2) 在【使用协方差矩阵】栏中选择分析时使用的协方差矩阵。

- 【在组内】。指定使用合并组内协方差矩阵进行分析。
- 【分组】。指定使用各组协方差矩阵进行分析。

(3) 在【输出】栏中选择生成到输出窗口中的分类结果。

- 【个案结果】。对每个观测输出判别分数、实际类、预测类(根据判别函数求得的分类结果)和后验概率等。选择此项，还可以选择其附属选项【将个案限制在前】，并在后面的小矩形框中输入观测数  $n$ ，含义为仅输出前  $n$  个观测的分类结果。观测数量大时可以选择此项。
- 【摘要表】。输出分类小结。给出正确分类观测数，即原始类和根据判别函数计算的预测类相同的观测数、错分观测(原始类和按判别函数计算出的类不同的观测)数和错分率。
- 【不考虑该个案时的分类】。输出每个观测的分类结果，所依据的判别函数是由除该观测以外的其他观测导出的，因此也称为交互校验结果。

(4) 在【图】栏中选择要求输出的统计图。可以同时选择几种输出的统计图形。

- 【合并组】。生成一张包括各类的散点图。该散点图是根据前两个判别函数值作出的。如果只有一个判别函数，就输出直方图。
- 【分组】。根据前两个判别函数值对每一类生成一张散点图，共分为几类就生成几张散点图。如果只有一个判别函数，就输出直方图。
- 【区域图】。根据函数值生成把观测分到各组中去的区域图。此种统计图把一张图的平面划分出与类数相同的区域，每一类占据一个区，各类的均值在各区中用星号标出。如果仅有一个判别函数，则不作此图。

(5) 选中对话框最下边的【使用均值替换缺失值】项，则观测某变量值缺失时，用该变量的均值代替缺失值。

## 10. 指定生成并保存在数据文件中的新变量

判别分析过程可以在数据文件中建立新变量。在主对话框中单击【保存】按钮，打开如图 13-41 所示的【判别分析：保存】对话框。

① 【预测组成员】。要求建立一个新变量，其值是根据判别分数、按后验概率最大预测的分类。每运行一次判别分析过程，就建立一个表明使用判别函数预测的观测属于哪一类的新变量。第一次运行建立新变量的变量名为 `dis_1`，如果在工作数据文件中不把前一次建立的新变量删除，则第  $n$  次运行建立的新变量默认的变量名为 `disn`。

② 【判别得分】。建立表明判别分数的新变量。该分数是由未标准化的判别系数乘以自变量的值，将这些乘积求和后加上常数得来的。每次运行判别分析过程都给出一组表明判别分数的新变量。建立几个判别函数就有几个判别分数变量。参与分析的观测共分为  $m$  类，则建立  $m-1$  个典则判别函数，指定该选项，就可以生成  $m-1$  个表明判别分数的新变量。例如，原始数据观测共分为 3 类，建立两个典则判别函数。第一次运行判别过程建立的新变量名为 `dis1_1`、



图 13-41 【判别分析：保存】对话框

dis2\_1, 第二次运行判别过程建立的新变量名为 dis1\_2、dis2\_2, …依此类推, 分别表示代入第一和第二个判别函数所得到的判别分数。

③【组成员概率】。要求建立新变量表明观测属于某一类的概率。有  $m$  类, 对一个观测就会给出  $m$  个概率值, 因此建立  $m$  个新变量。例如, 原始和预测分类数是 3, 指定该选项, 在第一次运行判别过程后, 给出的表明分类概率的新变量名为 dis1\_2、dis2\_2、dis3\_2。

11. 在主对话框中单击【确定】按钮提交执行。

13.5.3 判别分析实例

【例 8】 本例是统计学常用的实例。关于 3 种鸢尾花的花瓣、花萼的长、宽数据。共收集了 3 种鸢尾花, 每种 50 个观测, 共 150 个观测的数据。数据文件为 data13-05。

在数据窗口中定义 5 个变量: slen(花萼长)、swid(花萼宽)、plen(花瓣长)、pwid(花瓣宽), 是表明观测(鸢尾花)特征的变量)、spno(分类号)。分类的值标签是: 1—刚毛鸢尾花(Setosa), 2—变色鸢尾花(Versicolor), 3—弗吉尼亚鸢尾花(Virginica), 输入这些变量的值。观测标识变量 no 是为核对方便设置的, 非分析所需要。

使用系统默认值进行分析, 操作步骤如下。

(1) 读取数据文件 data13-05, 按【分析→分类→判别】顺序单击菜单项, 打开【判别分析】主对话框。

(2) 在主对话框中进行以下操作:

- ① 在源变量栏里选择 slen、swid、plen、pwid, 并移到【自变量】框中, 作为自变量。
- ② 在左侧的变量栏里选择变量 spno, 并移到【分组变量】框中, 作为分类变量。单击【定义范围】按钮, 在弹出的对话框中, 输入变量 spno 的数值范围, 【最小值】框中输入“1”, 【最大值】框中输入“3”。

表 13-24 基本数据信息

分类		有效的 N (列表状态)	
		未加权的	已加权的
刚毛鸢尾花	花萼长	50	50.000
	花萼宽	50	50.000
	花瓣长	50	50.000
	花瓣宽	50	50.000
变色鸢尾花	花萼长	50	50.000
	花萼宽	50	50.000
	花瓣长	50	50.000
	花瓣宽	50	50.000
弗吉尼亚鸢尾花	花萼长	50	50.000
	花萼宽	50	50.000
	花瓣长	50	50.000
	花瓣宽	50	50.000
合计	花萼长	150	150.000
	花萼宽	150	150.000
	花瓣长	150	150.000
	花瓣宽	150	150.000

(3) 单击【确定】按钮, 提交系统执行。

(4) 输出结果见表 13-24~表 13-29。输出结果解释如下。

表 13-24 所示为基本数据信息: 按变量 spno 确定分组、变量 spno 的标签是分类。下面同列单元格中是表明分类的 spno 变量的 3 个值标签。可以看出, 总共处理了 150 个(未加权)的观测。每类中各变量都有 50 个未加权的观测, 以下的分析中将使用 150 个(未加权)的观测。分组的观测数据表中的数据表明 spno = 1 为刚毛鸢尾花, spno = 2 为变色鸢尾花, spno = 3 为弗吉尼亚鸢尾花, 各有 50 个观测。3 种鸢尾花的每个观测的权重均为 1, 总权重均为 50。共有 150 个观测, 总权重为 150。

表 13-25 给出了典则判别函数特征值。给出的统计量自左至右有: 函数, 下面的单元格中的数字是函数代号; 特征值,

用于分析的前两个典则判别函数的特征值, 是组间平方和与组内平方和之比值; 最大特征值与组均值最大的向量对应, 第二大特征值对应着次大的组均值向量; 方差的%, 是方差的百分比; 累计%, 是累计百分比, 方差累计百分比最后累计值是 100%; 正则相关性, 是典则相关系数, 是组间平方和与总平方和之比的平方根。被平方的是由组间差异解释的总变异的比值。

表 13-26 所示为 Wilks' Lambda 统计量。该统计量进行检验的零假设是各组各变量均数相等。由于  $p < 0.001$ , 因此该判别函数能将两类很好地区分开。表中自左至右各列分别为: 比较

的函数编号；Wilks 的 Lambda 统计量值(也有称 U 统计量)值范围 0~1，越大表示组均值差异越小，值为 1 各组均值相等；卡方，是对 Wilks' Lambda 的卡方转换，用于确定其显著性；df 用于计算显著性水平的自由度；最后一列中两个函数的 Sig.都很小，说明建立的判别函数在统计上具有的显著性意义。

表 13-27 所示是标准化典则判别函数系数表。由此表可以看出，使用变量标签的两个判别函数分别如下。为分析方便，判别函数中使用的不是原变量名，而是变量标签：

$$y_1 = -0.346 \times \text{花萼长} - 0.525 \times \text{花萼宽} + 0.846 \times \text{花瓣长} + 0.613 \times \text{花瓣宽}$$
$$y_2 = 0.039 \times \text{花萼长} + 0.742 \times \text{花萼宽} - 0.386 \times \text{花瓣长} + 0.555 \times \text{花瓣宽}$$

注意：上述是标准化的典则判别函数，若要计算标准化典则判别函数值(即标准化典则判别分数)，代入上述函数的自变量值必须是标准化以后的值。

表 13-28 所示为结构阵，即合并类内相关阵，是判别变量与标准化的典则判别函数之间的相关。变量按函数内相关的绝对值大小排列，每个变量和任何一个判别函数之间相关系数绝对值最大的标有“\*”。

表 13-25 典则判别函数特征值表

函数	特征值	方差的 %	累积 %	正则相关性
1	30.419 <sup>a</sup>	99.0	99.0	.984
2	.293 <sup>a</sup>	1.0	100.0	.476

a. 分析中使用了前 2 个典型判别式函数。

表 13-26 判别函数的有效性检验

Wilks 的 Lambda				
函数检验	Wilks 的 Lambda	卡方	df	Sig.
1 到 2	.025	538.950	8	.000
2	.774	37.351	3	.000

表 13-27 标准典则判别函数的系数

	函数	
	1	2
花萼长	-.346	.039
花萼宽	-.525	.742
花瓣长	.846	-.386
花瓣宽	.613	.555

表 13-28 结构矩阵

	函数	
	1	2
花瓣长	.726*	.165
花萼宽	-.121	.879*
花瓣宽	.651	.718*
花萼长	.221	.340*

判别变量和标准化典型判别式函数之间的汇聚组间相关性  
按函数内相关性的绝对大小排序的变量。

\*. 每个变量和任意判别式函数间最大的绝对相关性

表 13-29 所示为类均值(重心)处的典则判别函数值。刚毛鸢尾花类中心的函数值为  $y_1 = -7.392$ ， $y_2 = 0.219$ ；变色鸢尾花类中心的函数值为  $y_1 = 1.763$ ， $y_2 = -0.737$ ；弗吉尼亚鸢尾花类中心的函数值为  $y_1 = 5.629$ ， $y_2 = 0.518$ 。

未标准化的典则判别函数中心值在各变量均值处。

【例 9】 仍然使用【例 1】，说明选项的作用。

操作步骤如下：

- (1) 再次打开【判别分析】主对话框，选择分析变量和分类变量的操作同【例 8】。
- (2) 在主对话框中，单击【分类】按钮，打开【判别分析：分类】对话框，选择分类参数。
  - ① 在【先验概率】栏中选择【所有组相等】项。
  - ② 在【使用协方差矩阵】栏中选择【在组内】项。

表 13-29 类中心

分类	函数	
	1	2
刚毛鸢尾花	-7.392	.219
变色鸢尾花	1.763	-.737
弗吉尼亚鸢尾花	5.629	.518

在组均值处评估的非标准化典型判别式函数

③ 在【图】栏中选择输出的统计图。选择【合并组】，要求作综合散点图；选择【分组】，要对每类作一个散点图；选择【区域图】，要求作按类分观测的区域图。

④ 在【输出】栏中选择【摘要表】，要求输出有关分类的数据。

表 13-30 原始数据的描述统计量

分类		均值	标准差	有效的 N (列表状态)	
				未加权的	已加权的
刚毛鸢尾花	花萼长	50.06	3.525	50	50.000
	花萼宽	34.28	3.791	50	50.000
	花瓣长	14.62	1.737	50	50.000
	花瓣宽	2.46	1.054	50	50.000
变色鸢尾花	花萼长	59.36	5.162	50	50.000
	花萼宽	27.66	3.147	50	50.000
	花瓣长	42.60	4.699	50	50.000
	花瓣宽	13.26	1.978	50	50.000
弗吉尼亚鸢尾花	花萼长	66.38	7.128	50	50.000
	花萼宽	29.82	3.218	50	50.000
	花瓣长	55.60	5.540	50	50.000
	花瓣宽	20.26	2.747	50	50.000
合计	花萼长	58.60	8.633	150	150.000
	花萼宽	30.59	4.363	150	150.000
	花瓣长	37.61	17.682	150	150.000
	花瓣宽	11.99	7.622	150	150.000

显示合并类内协方差矩阵；选择【分组协方差】，要求显示各类的协方差矩阵；选择【总体协方差】，要求显示总协方差矩阵。

(4) 在主对话框中单击【保存】按钮，打开【判别分析：保存】对话框，选择要求保存在工作数据文件中的新变量。选择【预测组成员】，要求建立表明预测的类成员号的新变量；选择【判别得分】，要求建立表明判别分数的新变量；选择【组成员概率】，要求建立表明观测作为各组成员的概率。

(5) 两点说明：

① 由于在主对话框中仍然是选择了【一起输入自变量】项，因此不能对判别分析方法进行进一步的选择。分析变量为所有 4 个反映鸢尾花特点的关于花瓣长、宽以及花萼长宽的变量，一起进入判别函数。

② 使用所有观测进行判别分析，同时因为工作数据文件中没有一个表示选择观测的变量，因此无须也不能对观测进行进一步的选择。

输出结果见表 13-30~表 13-37 和图 13-42、图 13-43。结果与【例 8】相同的表格不再重复列出。分析的输出结果解释如下：

表 13-30 所示是原始数据描述统计量，除包括基本数据信息外，还有各类中各变量的均值、标准差和整个样本的总均值、总标准差。

表 13-31 所示是各组均值相等的检验结果。进行的检验假设：各类中同变量均值相等。如果假设成立，说明根据各判别变量所作的原始分类是没有实际意义的。要么是分类错误，要么是选作判别的自变量不能充分显示分类特征。无论什么原因，进一步的输出结果分析均是无意义的。

如果有的变量方差分析结果表明变量对判别分析有意义，有的变量对判别分析无意义，则 要改变判别分析方法，以便自动剔除对判别分析无意义的变量。

如果拒绝假设，说明原始分类有意义。同时，可以认为判别自变量能够表明分类特征。本

(3) 在主对话框中，单击【统计量】按钮，打开相应对话框，选择要求输出的统计量。

① 在【描述性】栏中选择要输出的统计量，选择【均值】、【标准差】。选择【单变量 ANOVA】，要求输出每个变量的方差分析结果。检验的假设是：各类中同一自变量均值都相等。

② 在【函数系数】栏中选择判别函数系数。选择【Fisher】要求输出费雪系数；选择【未标准化】，要求输出未标准化的判别函数的系数。

③ 在【矩阵】栏中选择要输出的矩阵。选择【组内相关】，要求显示合并类内相关矩阵；选择【组内协方差】，要求



例的方差分析结果 Sig. 值均小于 0.001，说明 4 个判别变量都能很好地体现分类特征。但这并不说明所有变量相互独立，都应该出现在判别函数中。

表 13-32 所示是合并类内相关矩阵和合并类内协方差矩阵。合并类内协方差阵各元素的值是各类协方差阵相应元素值之平均值。合并类内相关阵各元素的值是各类相关阵相应元素值之平均值。由此表可以看出，花瓣长和花萼长之间的协方差值 16.129 和相关系数值 0.683 比较大，这就可以提出一个问题：是否它们之间不独立，在求出的判别函数中可否剔除一个变量呢？

表 13-31 各组均值相等的检验

	Wilks 的 Lambda	F	df1	df2	Sig.
花萼长	.397	111.847	2	147	.000
花萼宽	.598	49.371	2	147	.000
花瓣长	.059	1179.052	2	147	.000
花瓣宽	.071	960.007	2	147	.000

表 13-32 合并类内相关阵和协方差阵

		花萼长	花萼宽	花瓣长	花瓣宽
协方差	花萼长	29.960	8.767	16.129	4.340
	花萼宽	8.767	11.542	5.033	3.145
	花瓣长	16.129	5.033	18.597	4.287
	花瓣宽	4.340	3.145	4.287	4.188
相关性	花萼长	1.000	.471	.683	.387
	花萼宽	.471	1.000	.344	.452
	花瓣长	.683	.344	1.000	.486
	花瓣宽	.387	.452	.486	1.000

a. 协方差矩阵的自由度为 147。

表 13-33 所示是各类协方差阵和总协方差阵。除刚毛鸢尾花外，其余两种鸢尾花的协方差矩阵中协方差系数(除自协方差外)最大的是花瓣长和花萼长之间的协方差值，分别为 18.290 和 28.461，合计栏中的结果也一样为 130.036，因此有进一步分析的必要。

表 13-34 给出了未标准化的典则判别函数的系数。由于是未标准化的典则判别函数，因此有常数项，从表中可以得出两个判别函数分别为：

$y_1 = -0.063 \times \text{花萼长} - 0.155 \times \text{花萼宽} + 0.196 \times \text{花瓣长} + 0.299 \times \text{花瓣宽} - 2.526$

$y_2 = 0.007 \times \text{花萼长} + 0.218 \times \text{花萼宽} - 0.089 \times \text{花瓣长} + 0.271 \times \text{花瓣宽} - 6.987$

根据这两个典则判别函数可以计算出判别分数，根据各观测的两个判别分数可以画出区域图或散点图。

表 13-33 各类协方差阵和总协方差阵

协方差矩阵 <sup>a</sup>					
分类		花萼长	花萼宽	花瓣长	花瓣宽
刚毛鸢尾花	花萼长	12.425	9.922	1.636	1.033
	花萼宽	9.922	14.369	1.170	.930
	花瓣长	1.636	1.170	3.016	.607
	花瓣宽	1.033	.930	.607	1.111
变色鸢尾花	花萼长	26.643	8.288	18.290	5.578
	花萼宽	8.288	9.902	8.127	4.049
	花瓣长	18.290	8.127	22.082	7.310
	花瓣宽	5.578	4.049	7.310	3.911
弗吉尼亚鸢尾花	花萼长	50.812	8.090	28.461	6.409
	花萼宽	8.090	10.355	5.804	4.456
	花瓣长	28.461	5.804	30.694	4.943
	花瓣宽	6.409	4.456	4.943	7.543
合计	花萼长	74.537	-4.683	130.036	53.507
	花萼宽	-4.683	19.036	-33.056	-12.083
	花瓣长	130.036	-33.056	312.670	129.803
	花瓣宽	53.507	-12.083	129.803	58.101

a. 总的协方差矩阵的自由度为 149。

表 13-34 典则判别函数的系数

	函数	
	1	2
花萼长	-.063	.007
花萼宽	-.155	.218
花瓣长	.196	-.089
花瓣宽	.299	.271
(常量)	-2.526	-6.987

非标准化系数

有关判别函数的信息共输出 5 个表：典则判别函数特征值表、有效性检验表、类中心的函

数值表,表明自变量与函数之间相关的结构矩阵表和标准化、非标准化的典则判别函数系数表。前 5 个表分别与表 13-25、表 13-31 和表 13-29、表 13-28 相同,在此不再列出与说明。

表 13-35 所示是分析中使用的各类的先验概率。由于在【判别分析: 分类】对话框中选择的是【所有组相等】,因此各为 0.333,分析中使用的观测数加权、未加权都是 50。

表 13-36 所示是用判别函数对观测分类的结果。显示了费雪线性判别函数的系数。根据系数表可以总结出各类判别函数如下。

刚毛鸢尾花:  $F_1 = 1.687 \times \text{花萼长} + 2.695 \times \text{花萼宽} - 0.880 \times \text{花瓣长} - 2.284 \times \text{花瓣宽} - 80.268$

变色鸢尾花:  $F_2 = 1.101 \times \text{花萼长} + 1.070 \times \text{花萼宽} + 1.001 \times \text{花瓣长} + 0.197 \times \text{花瓣宽} - 71.196$

弗吉尼亚鸢尾花:  $F_3 = 0.865 \times \text{花萼长} + 0.747 \times \text{花萼宽} + 1.647 \times \text{花瓣长} + 1.695 \times \text{花瓣宽} - 103.896$

表 13-35 分析中的先验概率

分类	先验	用于分析的案例	
		未加权的	已加权的
刚毛鸢尾花	.333	50	50.000
变色鸢尾花	.333	50	50.000
弗吉尼亚鸢尾花	.333	50	50.000
合计	1.000	150	150.000

表 13-36 各类的分类函数的系数

	分类		
	刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花
花萼长	1.687	1.101	.865
花萼宽	2.695	1.070	.747
花瓣长	-.880	1.001	1.647
花瓣宽	-2.284	.197	1.695
(常量)	-80.268	-71.196	-103.890

Fisher 的线性判别式函数

使用费雪判别函数的方法是测得一种鸢尾花的 4 个自变量: 花萼长、花萼宽、花瓣长、花瓣宽的值,将 4 个自变量值代入上述 3 个函数式,得到 3 个函数值。比较这 3 个函数值,哪个值大就可以判断被测量的花属于哪类鸢尾花。例如,第一个观测: 花萼长 slen = 50,花萼宽 swid = 33,花瓣长 plen = 14,花瓣宽 pwid = 2。代入函数 1,得到  $F_1 = 75.751$ ;代入函数 2,得到  $F_2 = 33.572$ ;代入函数 3,得到  $F_3 = -9.547$ ;比较 3 个值,可以看出  $F_1 = 75.751$  最大,据此得出第一个观测属于刚毛鸢尾花。

表 13-37 所示是预测分类结果小结,是一个判别回代小结。所谓回代就是对一个被测试的观测使用下述方法判别属于的类:

- 使用除该观测以外的观测,求出线性判别函数。
- 使用求出的判别函数对这个观测进行判别得出该观测属于哪一类。
- 对每个观测均使用该方法进行判别,然后统计错判率。与原始数据中的 spno 变量值进行比较得出错判概率。

表 13-37 预测分类结果小结

		预测组成员			合计
		刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花	
初始	计数				
	刚毛鸢尾花	50	0	0	50
	变色鸢尾花	0	48	2	50
	弗吉尼亚鸢尾花	0	1	49	50
%	刚毛鸢尾花	100.0	.0	.0	100.0
	变色鸢尾花	.0	96.0	4.0	100.0
	弗吉尼亚鸢尾花	.0	2.0	98.0	100.0

a. 已对初始分组案例中的 98.0% 个进行了正确分类。

从表中可以看出,利用判别函数回代的结果,刚毛鸢尾花的错判率为 0%; 变色鸢尾花的错判率为 4%,是错判为第三类弗吉尼亚鸢尾花了; 弗吉尼亚鸢尾花的错判率为 2%,有一个观

测被错判为第二类变色鸢尾花了。回代结果有 98% 判别正确。

图 13-42 所示是区域图。横坐标用第一个典则变量，纵坐标用第二个典则变量。3 种鸢尾花的典则变量值把一个典则变量组成的坐标平面划分成 3 个区域。可以看出变色鸢尾花的数据居于另外两种鸢尾花数据之间。

图 13-42 下面的列表是区域图中的标记符号。分别用 1、2、3 表明刚毛鸢尾花、变色鸢尾花、弗吉尼亚鸢尾花的区域，用 “\*” 表明各类鸢尾花的数据重心。

中心坐标(典则判别函数 1 值，典则判别函数 2 值)表示 3 种鸢尾花的中心分别为刚毛鸢尾花中心(-7.392, 0.219)、变色鸢尾花中心(1.763, -0.737)、弗吉尼亚鸢尾花中心(5.629, 0.518)。中心数据见表 13-27。

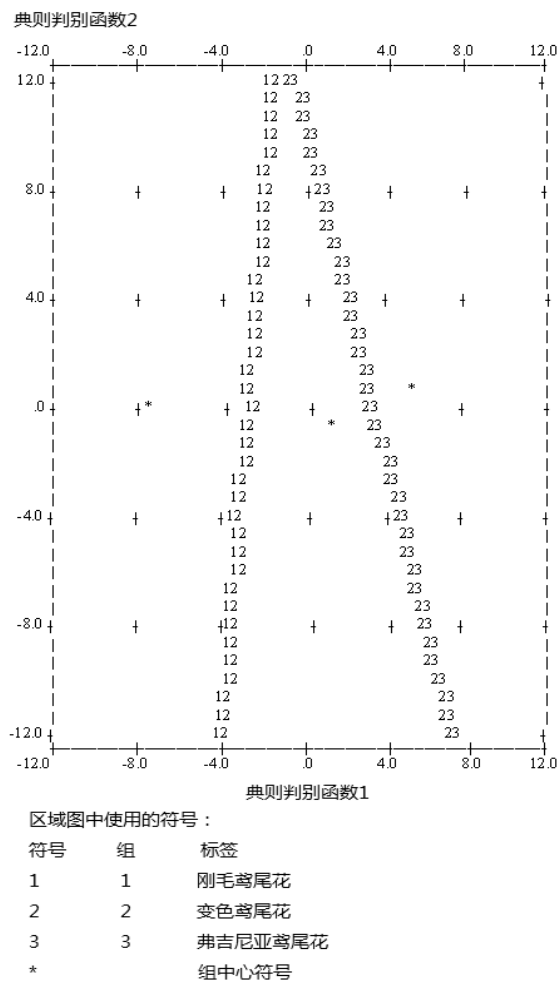


图 13-42 各类区域图及其标记说明

图 13-43 所示为每种鸢尾花的散点图，还有一个总的分类散点图。横坐标是典则判别函数 1，纵坐标是典则判别函数 2；根据自变量值计算两个典则判别函数值后作出；总图中可以看出各类之间的关系。

用 SAVE 子命令建立的新变量的信息表显示在输出文本的最开始处。新变量在数据窗中的情况见图 13-44。

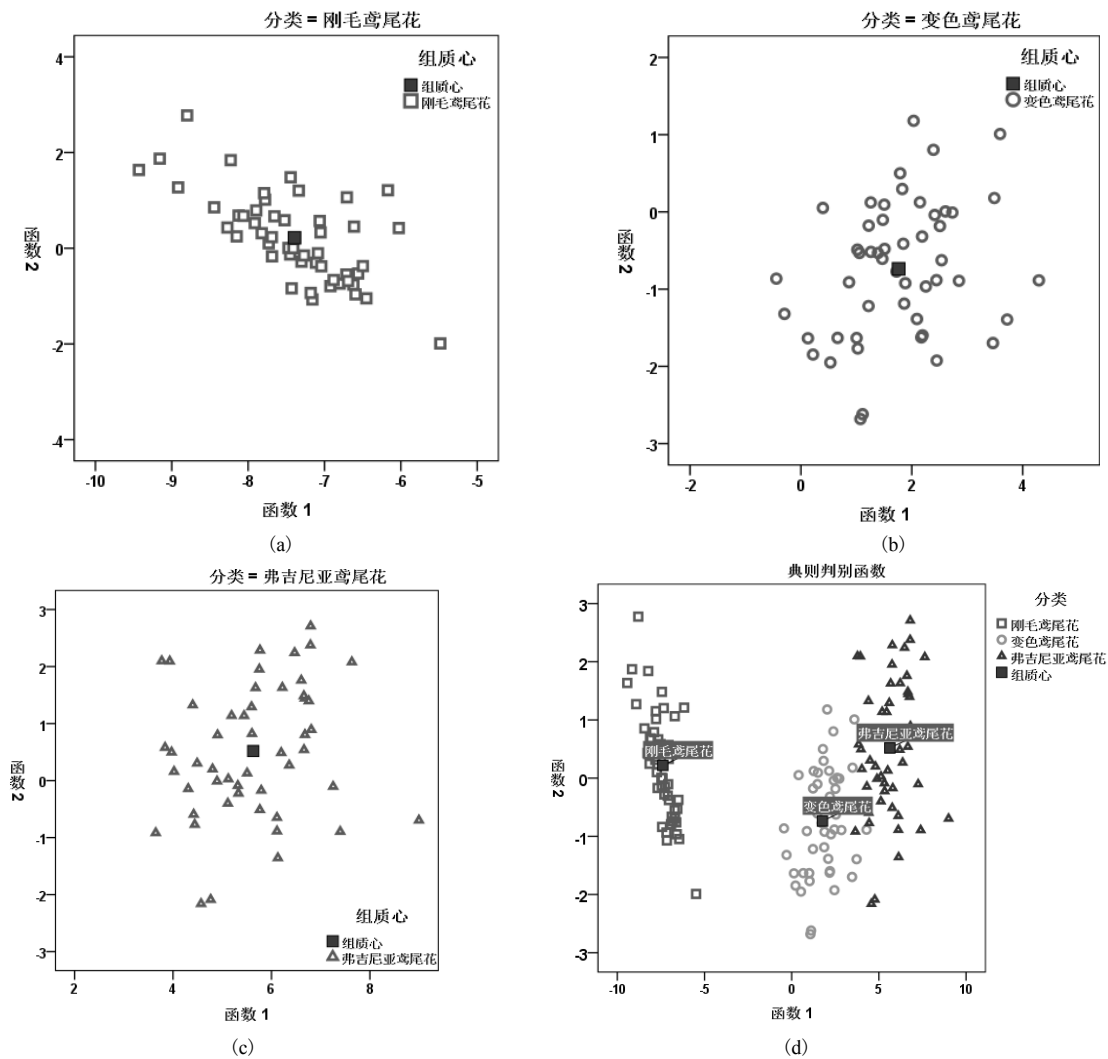


图 13-43 以典则判别函数为坐标的散点图

Figure 13-44 shows the Variable View of the IBM SPSS Statistics Data Editor for the file "data13-05鸢尾花数据.sav". The table below represents the data shown in the screenshot:

名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量标准
1 no	数值(N)	3	0	编号	无	无	3	右	名义(N)
2 slen	数值(N)	3	0	花瓣长	无	无	4	右	度量(S)
3 swid	数值(N)	3	0	花瓣宽	无	无	4	右	度量(S)
4 plen	数值(N)	3	0	花瓣长	无	无	4	右	度量(S)
5 pwid	数值(N)	3	0	花瓣宽	无	无	4	右	度量(S)
6 spno	数值(N)	1	0	分类	{1, 刚毛鸢尾...}	无	9	右	名义(N)
7 Dis_1	数值(N)	1	0	用于分析 1 的预测组	{1, 刚毛鸢尾...}	无	9	右	名义(N)
8 Dis_1_1	数值(N)	11	5	用于分析 1 的来自函数 1 的判别得分	无	无	7	右	度量(S)
9 Dis_2_1	数值(N)	11	5	用于分析 1 的来自函数 2 的判别得分	无	无	6	右	度量(S)
10 Dis_1_2	数值(N)	8	5	用于分析 1 的组 1 的成员概率	无	无	7	右	度量(S)
11 Dis_2_2	数值(N)	8	5	用于分析 1 的组 2 的成员概率	无	无	5	右	度量(S)
12 Dis_3_2	数值(N)	8	5	用于分析 1 的组 3 的成员概率	无	无	5	右	度量(S)
13									

图 13-44 由 SAVE 子命令建立的新变量信息表(变量视图)

新变量名采用系统默认方法。第一次运行命令程序建立的新变量的变量名及其数值含义说明

如下 (SPSS 20.0 软件自动生成的新变量的标签不能明确表示该变量的含义,因此再次再加以说明)。

- 变量 DIS\_1: 分析 1 预测的各观测所属类。
- 变量 DIS1\_1: 各观测在分析 1 中计算的未加权的典则变量 1 的值。
- 变量 DIS2\_1: 各观测在分析 1 中计算的未加权的典则变量 2 的值。
- 变量 DIS1\_2: 各观测在分析 1 中计算的属于第一类的概率。
- 变量 DIS2\_2: 各观测在分析 1 中计算的属于第二类的概率。
- 变量 DIS3\_2: 各观测在分析 1 中计算的属于第三类的概率。

如果分析方法不变,而且第一次运行产生的新变量没有从工作数据文件中删除,那么可以再运行一次判别分析过程,得到第二次分析生成的新变量。读者可以对比两次运行在工作数据文件中所列出的变量名,可以总结出系统默认变量名的规律。

图 13-45 所示是数据窗的数据视图。可以看到观测标号为 67、68 的记录,原始数据中这两个观测属于变色鸢尾花,但是根据判别函数预测的结果属于弗吉尼亚鸢尾花。这就是错分的两个观测,见表 13-37。

	no	slen	swid	plen	pwid	spno	Dis_1	Dis1_1	Dis2_1	Dis1_2	Dis2_2	Dis3_2	变量
60	31	56	25	39	11	变色鸢尾花	变色鸢尾花	1.00433	-1.63282	.00000	1.00000	.00000	
61	36	64	32	45	15	变色鸢尾花	变色鸢尾花	1.78949	.50070	.00000	.99865	.00135	
62	39	54	30	45	15	变色鸢尾花	变色鸢尾花	2.73144	-.00808	.00000	.97344	.02656	
63	47	67	31	44	14	变色鸢尾花	变色鸢尾花	1.25893	.12236	.00000	.99989	.00011	
64	50	57	26	35	10	变色鸢尾花	变色鸢尾花	-.29724	-1.32025	.00000	1.00000	.00000	
65	52	57	29	42	13	变色鸢尾花	变色鸢尾花	1.50938	-.47842	.00000	.99987	.00013	
66	53	65	26	46	15	变色鸢尾花	变色鸢尾花	2.85020	-.89216	.00000	.98597	.01403	
67	55	59	32	48	18	变色鸢尾花	弗吉尼亚鸢...	3.59207	1.00919	.00000	.26827	.73173	
68	56	60	27	51	16	变色鸢尾花	弗吉尼亚鸢...	4.29162	-.88613	.00000	.20951	.79049	
69	62	61	28	40	13	变色鸢尾花	变色鸢尾花	1.01881	-.48913	.00000	.99998	.00002	
70	63	55	24	38	11	变色鸢尾花	变色鸢尾花	1.02616	-1.76896	.00000	1.00000	.00000	
71	68	55	26	44	12	变色鸢尾花	变色鸢尾花	2.19268	-1.59802	.00000	.99954	.00046	

图 13-45 由 SAVE 子命令建立的新变量数据文件 (数据视图)

13.5.4 逐步判别分析与实例

1. 关于逐步判别分析

当研究某一事物分类时,往往对于哪些变量能够反映研究范围内事物的特性这一问题的认识还不够深刻,因此所选择的进行判别分析的变量不一定都能很好地反映类间差异。逐步判别分析假设已知的各类均属于多元正态分布,用逐步选择法选择最能反映类间差异的变量子集建立较好的判别函数。一个变量能否被选择为变量子集的成员进入模型主要取决于协方差分析的 F 检验的显著性水平。

逐步判别分析从模型中没有变量开始,每一步都对模型进行检测,把模型外对模型的判别力贡献最大的变量加入到模型中,同时考虑将已经在模型中但又不符合留在模型中的条件的变量从模型中剔除,直到模型中所有变量都符合留在模型中的判据、模型外的变量都不符合进入模型的判据时为止。

实际工作中应该把使用逐步判别分析选择变量的结果与在实践中对变量的认识相结合,会得到很好的判别分析模型。

## 2. 逐步判别分析方法与判据的选择

逐步判别的操作步骤参见第 13.5.2 节相关内容。在判别分析主对话框中应该选择【使用步进式方法】项。单击【方法】按钮,打开【判别分析:步进法】对话框,在【方法】栏进一步选择分析方法,在【标准】栏选择判据。

单击【方法】按钮,打开【判别分析:步进法】对话框,见图 13-38。系统默认的逐步判别方法是【Wilks' Lambda】。其判据是:进入模型的  $F$  值为  $F \geq 3.84$ ;从模型中剔除变量的判据是  $F$  值为  $F \leq 2.71$ 。不熟悉统计分析的读者可以不再进一步选择,直接使用系统默认的分析方法和判据。逐步判别方法和判据的选择以及要显示的输出内容均参见第 13.5.2 节相关内容。

**【例 10】** 为了容易比较,仍用鸢尾花的数据(数据文件 data13-05)作为逐步判别分析的数据。

**【例 8】、【例 9】** 中的程序都是使用全部变量建立判别函数。能否减少变量仍然得到较好的判别函数呢?采用 Wilks' Lambda 方法进行逐步判别分析。使用  $F$  值作为判据统计量。当  $F \geq 30$  时,变量进入模型;当  $F \leq 5$  时,变量从模型中移出。

(1) 操作步骤如下:

① 再次打开【判别分析】主对话框。

② 仍把全部 4 个自变量送入【自变量】框中,变量 spno 作为分类变量移到【分组变量】框中。单击【定义范围】按钮,在【判别分析:定义范围】对话框中,输入变量 spno 的数值范围:最小值“1”和最大值“3”。

③ 在主对话框中,选择【使用步进式方法】项,单击【方法】按钮,打开【判别分析:步进法】对话框。在【方法】栏中选择【Wilks' Lambda】项。在【标准】栏中选择【使用  $F$  值】项,并在【进入】框中输入“30”,【删除】框中输入“5”。在输出栏中选择【步进摘要】项,要求显示逐步选择变量子集的小结;选择【两两组间距离的  $F$  值】项,要求显示每两类之间的成对的  $F$  矩阵。

④ 在主对话框中单击【统计量】按钮,打开【判别分析:统计量】对话框。在【描述性】栏中选择【均值】、【单变量 ANOVA】项;在【函数系数】栏中选择【Fisher】、【未标准化】,在【矩阵】栏内选择【组内相关】项。

⑤ 在主对话框中单击【分类】按钮,打开【判别分析:分类】对话框。在【先验概率】栏内选择【所有组相等】项,即各组先验概率相等;在【使用协方差矩阵】栏内选择【在组内】项,使用组内协方差矩阵;在【输出】栏中选择【摘要表】,要求显示聚类回代结果的小结表。

⑥ 在主对话框中单击【保存】按钮,打开【判别分析:保存】对话框。选择【预测组成员】项,生成预测的观测所属类别的新变量;选择【判别得分】项,生成判别函数的分数新变量;选择【组成员概率】项,生成各观测属于各类的概率的新变量。

⑦ 在主对话框中单击【确定】按钮,提交运行。

(2) 输出结果见表 13-38~表 13-50。与【例 8】重复的不再列出。

(3) 输出结果解释。

① 没有列出的原始变量的描述统计量表与表 13-30 相同,是各类自变量的均值与标准差与自变量的总的均值和标准差。

从各类均值与标准差的比较中可以看出,各类鸢尾花中,变量花萼宽 swid 标准差值比其他变量值集中,总标准差最小。由此看到一个可能性:如果能从判别函数中减掉一个变量,这

个变量可能是花萼宽 swid，但与花瓣宽、花萼长的标准差在一个数量级上。因此还是要经过逐步判别分析才能最后确定。

② 表 13-38 所示为逐步判别分析前相关阵。从表 13-32 中虽然可以看出 4 个自变量都对区分鸢尾花的种类是有效变量，但根据表 13-38 相关矩阵可以看出，花瓣长和花萼长相关系数比较大，为 0.683。能否在判别函数中省掉一个自变量呢？这个问题由逐步判别分析来解决。

表 13-38 逐步判别前的自变量相关阵

	花萼长	花萼宽	花瓣长	花瓣宽
相关性	花萼长	花萼宽	花瓣长	花瓣宽
	1.000	.471	.683	.387
	.471	1.000	.344	.452
	.683	.344	1.000	.486
	.387	.452	.486	1.000

③ 表 13-39 所示是逐步判别分析的一个小结。“精确 F”栏内的统计量是一个  $F$  值，是该变量的均方与误差均方的比值。该值越大，Sig. 值越小，因此该值最大的先进入判别函数。当 Sig. 小于 0.05 或 0.01 时，拒绝零假设。显著性检验结果 Sig.=0.000，即小于 0.001，可以说这 3 个变量对判别的贡献都很显著。总之，说明该变量在不同类中均值不同是由于类间差异，而不是由随机误差引起的，即该变量在各组中均值差异显著。可以看出 3 个变量的  $F$  统计量都在 30 以上，这是选择进入判别函数的判据。

表 13-39 逐步判别分析小结

Wilks 的 Lambda									
步骤	变量数目	Lambda	df1	df2	df3	精确 F			
						统计量	df1	df2	Sig.
1	1	.059	1	2	147	1179.052	2	147.000	.000
2	2	.038	2	2	147	301.876	4	292.000	.000
3	3	.026	3	2	147	251.164	6	290.000	.000

④ 表 13-40、表 13-41 所示是根据 Wilks’ Lambda 值进行逐步选择变量并进行  $F$  检验的过程数据。每一步都计算该变量进入模型使 Wilks’ Lambda 值降低了多少，都是那个使总的 Wilks’ Lambda 值最小的变量进入判别函数。从这两个表可以看到逐步判别的每一步过程。判别分析在一个自变量进入模型后，对模型内各变量进行方差分析，在模型外的自变量进行方差分析和  $F$  检验。模型内的  $F$  检验  $F$  值小于 5 的自变量还要从模型中移出；模型外的自变量若  $F$  值大于 30，可以进入模型。

表 13-40 逐步进入模型的变量方差分析结果

步骤	容差	要删除的 F	Wilks 的 Lambda
1	花萼长	1.000	1179.052
2	花萼长	.882	1078.565
	花萼宽	.882	39.965
3	花萼长	.745	36.018
	花萼宽	.775	49.885
	花瓣宽	.672	33.060

表 13-41 各步模型外的变量方差分析结果

步骤	容差	最小容差	要输入的 F	Wilks 的 Lambda
0	花萼长	1.000	111.847	.397
	花萼宽	1.000	49.371	.598
	花瓣长	1.000	1179.052	.059
	花瓣宽	1.000	960.007	.071
1	花萼长	.533	.533	21.768
	花萼宽	.882	.882	39.965
	花瓣宽	.764	.764	24.435
2	花萼长	.470	.470	6.733
	花瓣宽	.672	.672	33.060
3	花萼长	.469	.469	4.159

- 表 13-41 的步骤 0，表明花瓣长  $F$  值最大， $F=1179.052$ ，Willk’s Lambda = 0.059 值最小，第一个进入模型的是花瓣长。
- 进入模型后，因为只有 1 个变量，在表 13-40 的步骤 1 中可以看出花瓣长第一个进入模型。
- 表 13-41 的步骤 1，是花瓣长进入模型后模型外的 3 个自变量的方差分析。花萼宽的  $F$  值最大，为 39.965，大于 30，Willk’s Lambda = 0.038 值最小，因此第二个进入模型的是花萼宽。

- 表 13-40 的步骤 2 中，花萼宽进入模型，模型内方差分析结果：花瓣长  $F = 1078.565$ ，花萼宽  $F = 39.965$ ，都大于 5，因此两个变量都保持在模型中。
- 表 13-41 的步骤 2 是模型外的变量方差分析结果， $F$  值最大的是花瓣宽， $F = 33.060$ ，也大于 30。Willk's Lambda = 0.026 值最小，应该自变量花瓣宽进入模型。
- 同样，表 13-40 的步骤 3 是花瓣宽进入模型后的方差分析结果， $F$  值均大于 5，3 个自变量仍保持在模型中。
- 表 13-41 的步骤 3 是模型外自变量方差分析，花萼长  $F = 4.159$ ，小于 30，该自变量不再进入判别函数模型。

模型外、内变量无进、无出，逐步判别分析的自变量选择结束。

⑤ 表 13-42 所示是在逐步判别分析过程的每一步中，在任意两类之间进行的方差分析，想看看在这一步选入模型中的自变量对任意两类之间的区分是否有效。 $F$  值越大，Sig 值越小，区分效果越好。行类与列类之间的方差分析结果显示在行列交叉单元格中。从表中可以看出，各步所选择的变量对任意两类的区分都是有效的。

输出窗中，在典型判别式函数摘要标题下的表格说明了使用选择的自变量导出的典则判别函数的结果。

⑥ 表 13-43 所示为两个典则判别函数的特征值表。可以看出与全模型的特征值相差不多。第一个函数仍占了总方差的 99%。

⑦ 表 13-44 所示为 Wilks' Lambda 值的卡方转换及检验。

⑧ 表 13-45 所示为标准化的典则判别函数系数表。可从中总结出标准化的典则判别函数。

$$y_1 = -0.640 \times \text{花萼宽} + 0.656 \times \text{花瓣长} + 0.642 \times \text{花瓣宽}$$
$$y_2 = 0.758 \times \text{花萼宽} - 0.367 \times \text{花瓣长} + 0.549 \times \text{花瓣宽}$$

注意：若用标准化的判别函数计算标准化的判别分数，必须代入标准化的自变量值。

⑨ 表 13-46 所示是判别自变量与标准化的典则判别函数之间的相关矩阵。标有字母“b”的自变量没有在判别函数中。

⑩ 表 13-47 所示为未标准化的典则判别函数。

$$y_1 = -0.188 \times \text{花萼宽} + 0.152 \times \text{花瓣长} + 0.314 \times \text{花瓣宽} - 3.715$$
$$y_2 = 0.223 \times \text{花萼宽} - 0.085 \times \text{花瓣长} + 0.268 \times \text{花瓣宽} - 6.842$$

使用未标准化的典则判别函数计算判别分数，要使用原始自变量值进行计算。

⑪ 表 13-48 所示为各类中心处的未标准化的典则判别函数值。与区域图中的“\*”坐标值对应。在区域图中以两个典则判别函数值为两个坐标轴，平面上的点表示为  $(f_1, f_2)$ ，则各种鸢尾花的类中心坐标分别为：刚毛鸢尾花  $(-7.180, 0.219)$ ，变色鸢尾花  $(1.708, -0.737)$ ，弗吉尼亚鸢尾花  $(5.472, 0.518)$ 。

输出窗中，分类统计量标题下的表格是使用判别函数对原始数据进行分类的结果数据。

⑫ 表 13-49 所示是使用逐步判别选择的变量进行线性判别分析结果。逐步判别选择变量的目的仍是要使用选择出的较少的自变量，导出判别函数，对观测进行进一步的判别，然后分析该判别函数的优劣。表 13-49 所示是费雪线性判别函数系数表。3 个线性判别函数如下：

$$F_1 = 3.452 \times \text{花萼宽} + 0.411 \times \text{花瓣长} - 2.425 \times \text{花瓣宽} - 60.280$$
$$F_2 = 1.564 \times \text{花萼宽} + 1.843 \times \text{花瓣长} + 0.105 \times \text{花瓣宽} - 62.685$$
$$F_3 = 1.135 \times \text{花萼宽} + 2.309 \times \text{花瓣长} + 1.622 \times \text{花瓣宽} - 98.632$$



使用线性判别函数，应代入各判别变量原始观测值，计算判别函数值即判别分数。

表 13-42 每步的类间比较

步骤	分类		刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花
1	刚毛鸢尾花	F		1052.420	2257.552
		Sig.		.000	.000
	变色鸢尾花	F	1052.420		227.185
		Sig.	.000		.000
	弗吉尼亚鸢尾花	F	2257.552	227.185	
		Sig.	.000	.000	
2	刚毛鸢尾花	F		768.305	1416.055
		Sig.		.000	.000
	变色鸢尾花	F	768.305		115.071
		Sig.	.000		.000
	弗吉尼亚鸢尾花	F	1416.055	115.071	
		Sig.	.000	.000	
3	刚毛鸢尾花	F		656.739	1316.404
		Sig.		.000	.000
	变色鸢尾花	F	656.739		129.425
		Sig.	.000		.000
	弗吉尼亚鸢尾花	F	1316.404	129.425	
		Sig.	.000	.000	

a. 步骤 1 的 1、147 自由度。  
b. 步骤 2 的 2、146 自由度。  
c. 步骤 3 的 3、145 自由度。

表 13-43 典则判别函数的特征值表

函数	特征值	方差的 %	累积 %	正则相关性
1	28.708 <sup>a</sup>	99.0	99.0	.983
2	.292 <sup>a</sup>	1.0	100.0	.476

a. 分析中使用了前 2 个典型判别式函数。

表 13-44  $\lambda$  值的卡方转换及卡方检验

Wilks 的 Lambda				
函数检验	Wilks 的 Lambda	卡方	df	Sig.
1 到 2	.026	532.603	6	.000
2	.774	37.454	2	.000

表 13-45 标准化的典则判别函数系数表

	函数	
	1	2
花萼宽	-.640	.758
花瓣长	.656	-.367
花瓣宽	.642	.549

表 13-46 结构矩阵

	函数	
	1	2
花瓣长	.747 <sup>*</sup>	.160
花萼长 <sup>b</sup>	.395 <sup>*</sup>	.319
花萼宽	-.125	.880 <sup>*</sup>
花瓣宽	.671	.714 <sup>*</sup>

判别变量和标准化典型判别式函数之间的汇聚组间相关性  
按函数内相关性的绝对大小排序的变量。

\*. 每个变量和任意判别式函数间最大的绝对相关性  
b. 该变量不在分析中使用。

表 13-47 未标准化的典则判别函数系数表

	函数	
	1	2
花萼宽	-.188	.223
花瓣长	.152	-.085
花瓣宽	.314	.268
(常量)	-3.715	-6.842

非标准化系数

表 13-48 各类中心的未标准化的判别函数值表

分类	函数	
	1	2
刚毛鸢尾花	-7.180	.219
变色鸢尾花	1.708	-.737
弗吉尼亚鸢尾花	5.472	.518

在组均值处评估的非标准化典型判别式函数

表 13-49 逐步判别选择的变量  
进行线性判别分析结果

	分类		
	刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花
花萼宽	3.452	1.564	1.135
花瓣长	.411	1.843	2.309
花瓣宽	-2.425	.105	1.622
(常量)	-60.280	-62.685	-98.632

Fisher 的线性判别式函数

⑬ 表 13-50 所示为逐步判别回代小结。

从表中数据可以看出，该表是用只包含 3 个变量的判别函数进行分类的小结。可以看出，对刚毛鸢尾花的分类错判率为 0%；对变色鸢尾花的分类有 2 个观测错判为弗吉尼亚鸢尾花了，错判率为 4%；对弗吉尼亚鸢尾花错判了 1 个，错判率为 2%。总的判断正确率为 98% (错判率为 2%)。虽然比起全模型来少了一个自变量，但错判率没有改变。由此也说明逐步判别的结果可行。

表 13-50 逐步判别回代小结

分类			预测组成员			合计
			刚毛鸢尾花	变色鸢尾花	弗吉尼亚鸢尾花	
初始	计数	刚毛鸢尾花	50	0	0	50
		变色鸢尾花	0	48	2	50
		弗吉尼亚鸢尾花	0	1	49	50
%		刚毛鸢尾花	100.0	.0	.0	100.0
		变色鸢尾花	.0	96.0	4.0	100.0
		弗吉尼亚鸢尾花	.0	2.0	98.0	100.0

a. 已对初始分组案例中的 98.0% 个进行了正确分类。

读者可以在【判别分析：分类】对话框中的【图】栏中选择【合并组】、【分组】和【区域图】分别作出分类散点图、各类散点总图和区域图，对各结果进行进一步的认识。由于篇幅关系，不再一一列出。

习 题 13

- 1. SPSS 提供几种聚类分析过程？各适合什么情况的聚类？
- 2. 聚类分析与判别分析对数据要求有什么不同？
- 3. 聚类分析之前一定要对变量进行标准化吗？为什么？
- 4. 变量聚类后如何根据聚类结果确定各类的代表变量？
- 5. 1976 年 74 个国家人口出生率和死亡率数据在数据文件 data13-06.xls 中。将数据转换成 SPSS 数据文件，以相同的主名保存成.sav 文件，根据出生率、死亡率聚类，绘制散点图。
- 6. 数据文件 data13-07.xls 的 sheet1 中是 28 名一级、25 名健将级标枪运动员测验的 6 项影响标枪成绩的项目成绩。据此求出判别运动员等级的判别函数；回代，给出错判率。sheet2 中是 14 名未知级别的运动员。运用判别函数对他们分类。转换成的 SPSS 数据文件请参考数据文件 data13-07 和 data13-07b。
- 7. 习题 6 中的 6 个与标枪成绩有关的项目彼此是否相关？能否进行变量聚类？并找出各类中有代表性的项目 (变量)。
- 8. 用逐步判别法再求判别函数，与用全部变量求出的判别函数比较错判率。

# 第 14 章 因子分析与对应分析

在各个领域的科学研究中，往往需要对反映事物的多个变量进行大量的观测，收集大量数据以便进行分析寻找规律。多变量大样本无疑会为科学研究提供丰富的信息，但也在一定程度上增加了数据采集的工作量，更重要的是在大多数情况下，由于许多变量之间可能相关，增加了问题分析的复杂性，同时对分析带来不便。如果分别分析每个指标，分析又可能是孤立的，而不是综合的。盲目减少指标会损失很多信息，容易产生错误的结论。因此需要找到一个合理的方法，减少分析指标的同时，尽量减少原指标包含信息的损失，对所收集的资料作全面的分析。由于各变量间存在一定的相关关系，因此有可能用较少的综合指标，分别综合存在于各变量中的各类信息。这就是降维。SPSS 收集的降维方法在【分析】的【降维】菜单中。【因子分析】、【对应分析】和【最优尺度】分析就是这样的降维方法，见图 14-1。

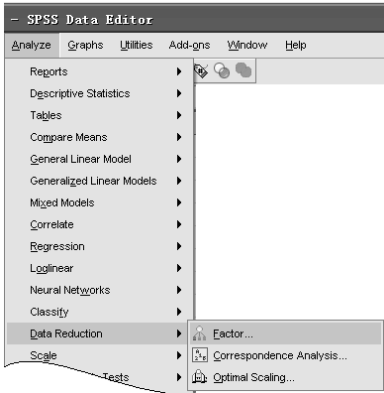


图 14-1 调用因子分析过程命令

## 14.1 主成分分析与因子分析

### 14.1.1 主成分分析与因子分析概述

#### 1. 主成分分析的概念

##### (1) 什么是主成分分析？

在各领域的科学研究中，为了全面客观地分析问题，往往要考虑从多方面观察所研究的对象，要收集多个观察指标的数据。如果一个一个地分析这些指标，无疑会造成对研究对象的片面认识，也不容易得出综合的、一致性很好的结论。主成分分析就是考虑各指标间的相互关系，利用降维的思想把多个指标转换成较少的几个互不相关的综合指标，从而使进一步研究变得简单的一种统计方法。

现举例说明主成分分析。儿童身高和体重两个变量之间的关系(见表 14-1)可以使用散点图表示，见图 14-2。显然，这两个变量之间存在线性关系。数据 $(h_i, w_i)$ 各点散布在一条直线周围，其中 $i=1\sim n$ 。

现在以该直线为坐标轴 $p_1$ ，以该轴的垂直线为另一个坐标轴 $p_2$ 。因为所有观测点均在坐标轴 $p_1$ 周围，而 $p_1$ 与 $p_2$ 是两个相互垂直的坐标轴，因此彼此不相关。

原观测点可以表示为 $(p_{1i}, p_{2i})$ ， $i=1\sim n$ 。可以认为， $n$ 个观测的差异主要表现在 $p_1$ 方向上，而在 $p_2$ 方向上差异很小。

由此得出结论，可以用一个指标 $p_1$ 来代替原始变量 $h, w$ 研究 $n$ 个观测对象的差异。 $p_1$ 、

$p_2$  可以用原始变量  $h$ 、 $w$  的线性组合来表示, 即

$$\begin{cases} p_1 = l_{11}h + l_{12}w \\ p_2 = l_{21}h + l_{22}w \end{cases}$$

式中, 系数  $l_{11}$ 、 $l_{12}$ 、 $l_{21}$ 、 $l_{22}$  是可以计算出来的。

表 14-1 身高体重数据

变量 观测 I	身高 $h$	体重 $w$
1	$h_1$	$w_1$
2	$h_2$	$w_2$
3	$h_3$	$w_3$
4	$h_4$	$w_4$
...	...	...
$n$	$h_n$	$w_n$

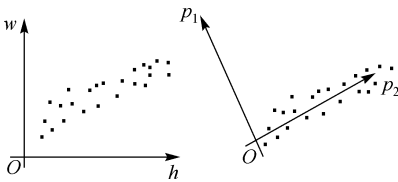


图 14-2 主成分概念示意图

如果  $p_1$  代表了观测值变化最大的方向(即沿该方向观测值方差最大), 而且  $p_2$  和  $p_1$  正交, 则称  $p_1$  为  $h$ 、 $w$  的第一主成分,  $p_2$  称为  $h$ 、 $w$  的第二主成分。这种分析方法称为主成分法。可以看出:

- ① 新变量  $p_1$ 、 $p_2$  是原始变量  $h$ 、 $w$  的线性函数。
- ②  $p_1$  与  $p_2$  相互垂直, 即两个新变量不相关。

由此推广到一般情况, 实测变量  $x_1 \sim x_m$ , 共测得  $n$  个观测, 数据如表 14-2 所示。

表 14-2 参与因子分析的观测与变量数据

变量 $j$ 观测 $i$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	...	$x_m$
1	$x_{11}$	$x_{12}$	$x_{13}$	$x_{14}$	$x_{15}$	...	$x_{1m}$
2	$x_{21}$	$x_{22}$	$x_{23}$	$x_{24}$	$x_{25}$	...	$x_{2m}$
3	$x_{31}$	$x_{32}$	$x_{33}$	$x_{34}$	$x_{35}$	...	$x_{3m}$
4	$x_{41}$	$x_{42}$	$x_{43}$	$x_{44}$	$x_{45}$	...	$x_{4m}$
5	$x_{51}$	$x_{52}$	$x_{53}$	$x_{54}$	$x_{55}$	...	$x_{5m}$
...	...	...	...	...	...	...	...
$n$	$x_{n1}$	$x_{n2}$	$x_{n3}$	$x_{n4}$	$x_{n5}$	...	$x_{nm}$

在原始变量的  $m$  维空间中, 找到新的  $m$  个坐标轴, 新变量与原始变量的关系可以表示为

$$\begin{cases} p_1 = l_{11}x_1 + l_{12}x_2 + l_{13}x_3 + \cdots + l_{1m}x_m \\ p_2 = l_{21}x_1 + l_{22}x_2 + l_{23}x_3 + \cdots + l_{2m}x_m \\ p_3 = l_{31}x_1 + l_{32}x_2 + l_{33}x_3 + \cdots + l_{3m}x_m \\ \vdots \\ p_m = l_{m1}x_1 + l_{m2}x_2 + l_{m3}x_3 + \cdots + l_{mm}x_m \end{cases}$$

这  $m$  个新变量中可以找到  $l$  个新变量( $l < m$ ) 能解释原始数据大部分方差所包含的信息, 包含的信息量是原始数据包含信息量的绝大部分; 其余  $m-l$  个新变量对方差影响很小。我们称这  $m$  个新变量为原始变量的主成分, 每个新变量均为原始变量的线性组合。

(2) 主成分分析中的统计量。

前面提到, 使得方差最大的  $l$  个互相正交的方向及沿这些方向的方差是一个特征值问题的特征向量和特征值。这些特征值和特征向量为特征方程  $Ax = \lambda x$  的解, 这里  $A$  为样本协方差阵或样本相关阵。如果用样本相关阵, 可以避免由于各变量量纲不同而产生的问题; 如果用样本

协方差阵, 应该对原始变量进行标准化, 这在 SPSS 中是自动完成的。主成分分析的主要统计量如表 14-3 所示。

表 14-3 主成分分析中的主要统计量

成分号 $i$	特征值 $\lambda_i$	贡献率 $\lambda_i/m$	累计贡献率	特征向量 $L_i$ : $l_{i1} l_{i2} \cdots l_{im}$
1	$\lambda_1$	$\lambda_1/m$	$\lambda_1/m$	$L_1$ : $l_{11} l_{12} \cdots l_{1m}$
2	$\lambda_2$	$\lambda_2/m$	$(\lambda_1 + \lambda_2)/m$	$L_2$ : $l_{21} l_{22} \cdots l_{2m}$
3	$\lambda_3$	$\lambda_3/m$	$(\lambda_1 + \lambda_2 + \lambda_3)/m$	$L_3$ : $l_{31} l_{32} \cdots l_{3m}$
...	...	...	...	...
$m$	$\lambda_m$	$\lambda_m/m$	$m$	$L_m$ : $l_{m1} l_{m2} \cdots l_{mm}$

① 特征方程的根, 通常用  $\lambda$  表示。有  $m$  个变量, 就有  $m$  个特征方程的根。它是确定主成分数目的根据。SPSS 软件输出列出的特征方程的根已经是经过重新排序重新命名的结果。最大的为  $\lambda_1$ , 最小的为  $\lambda_m$ 。

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_m$$

该统计量反映的是原始变量的总方差在各成分上重新分配的结果。

根据方差的定义, 第  $i$  个主成分的方差是总方差在各主成分上重新分配后, 在第  $i$  个成分上分配的结果, 在数值上等于第  $i$  个特征值, 即

$$S_{P_i} = \frac{\sum_{i=1}^m (p_i - \bar{p}_i)^2}{n-1} = \lambda_i$$

$\sum_{i=1}^m \lambda_i = m$  原始变量个数  $m$  等于特征值的数目  $m$ ,  $m$  个特征值之方差总和等于  $m$  个特征值之和。等于  $m$ , 即等于标准化的原始变量的方差之总和。

② 各成分之贡献率定义: 各成分所包含的信息占总信息的百分比。用方差衡量变量所包含的信息量, 则每个成分所提供方差占总方差 ( $m$ ) 的百分比即该成分的贡献率, 即  $P_i$  的贡献率为

$$\frac{\lambda_i}{\sum_{i=1}^m \lambda_i} = \frac{S_{P_i}}{\sum_{i=1}^m S_{P_i}} = \frac{\lambda_i}{m}$$

③ 前  $k$  个成分的累计贡献率为

$$\sum_{i=1}^k \frac{\lambda_i}{\sum_{i=1}^m \lambda_i} = \sum_{i=1}^k \frac{\lambda_i}{m}$$

通常取累计贡献率大于等于 80% 来确定取前  $k$  个成分作为该研究问题的主成分。

④ 确定取几个成分作为主成分的判定方法有两种:

- 取所有特征值大于 1 的成分作为主成分。
- 根据累计贡献率达到的百分比值确定。如取累计贡献率达到 80%, 其含义是此前  $l$  个成分 (新变量) 所包含的信息占原始变量包含的总信息的 80%, 其余  $m-l$  个新变量对方差影响很小, 如果认为可以接受, 则取前  $l$  个成分作为主成分。

⑤ 特征向量是各成分表达式中标准化原始变量的系数向量，就是各成分的特征向量。得出特征向量，就可以写出每个成分的表达式。注意，前面公式中得到的使  $S_{p_i} = \lambda_i$  的各个成分  $p_i$  的系数  $(l_{i1}, l_{i2}, \dots, l_{im})$  是单位特征向量，并不是 SPSS 输出中的成分矩阵中的系数。而成分矩阵中的各个分量的系数为此单位特征向量乘以相应的特征值的平方根的结果。如果令

$$a_{ij} = \sqrt{\lambda_i} l_{ij} \quad i, j = 1, \dots, m$$

那么， $a_{ij}$  为第  $i$  个成分和第  $j$  个变量的相关系数，也称为载荷。SPSS 中的成分图即载荷图选项就是由成分矩阵中各个分量系数点作出来的。

⑥ 主成分分数。根据主成分表达式和各观测中各变量值计算出的成分值，它与上面关于  $p_i$  公式中(用  $a_{ij}$  代替  $l_{ij}$ ，并且把变量  $x_j$  标准化之后)得到的  $p_i$  成比例，称为该观测的该成分的分数。该成分是第几个主成分，就称该值为第几个主成分分数。如果在输出中选了该项，则在原始数据中会增加对每个观测值所计算的主成分的分数。

2. 因子分析的概念

(1) 什么是因子分析。

探讨存在相关关系的变量之间，是否存在不能直接观察到但对可观测变量的变化起支配作用的潜在因子的分析方法称为因子分析。因子分析就是寻找潜在的起支配作用的因子模型的方法。

设有原始变量  $x_1, x_2, x_3, \dots, x_m$ 。它们与潜在因子之间的关系可以表示为

$$\begin{cases} x_1 = b_{11}z_1 + b_{12}z_2 + b_{13}z_3 + \dots + b_{1m}z_m + e_1 \\ x_2 = b_{21}z_1 + b_{22}z_2 + b_{23}z_3 + \dots + b_{2m}z_m + e_2 \\ x_3 = b_{31}z_1 + b_{32}z_2 + b_{33}z_3 + \dots + b_{3m}z_m + e_3 \\ \vdots \\ x_m = b_{m1}z_1 + b_{m2}z_2 + b_{m3}z_3 + \dots + b_{mm}z_m + e_m \end{cases}$$

式中， $z_1 \sim z_m$  为  $m$  个潜在因子，是各原始变量都包含的因子，称共性因子； $e_1 \sim e_m$  为  $m$  个只包含在某个原始变量之中的，只对一个原始变量起作用的个性因子，是各变量特有的特殊因子。

共性因子与特殊因子相互独立。找出共性因子是因子分析的主要目的。计算出结果后要对共性因子的实际含义进行探讨，并命名。

进行因子分析的方法很多，常用的是主成分法。如果特殊因子可以忽略，可以使用主成分分析的计算方法进行因子分析。

(2) 因子分析中的统计量。

① 因子与因子载荷。根据累计贡献率尽量大的原则决定公因子数。公因子数为  $k$ ，初始因子模型为

$$\begin{cases} x'_1 = \alpha_{11}f_1 + \alpha_{12}f_2 + \dots + \alpha_{1k}f_k + e_1 \\ x'_2 = \alpha_{21}f_1 + \alpha_{22}f_2 + \dots + \alpha_{2k}f_k + e_2 \\ x'_3 = \alpha_{31}f_1 + \alpha_{32}f_2 + \dots + \alpha_{3k}f_k + e_3 \\ \vdots \\ x'_m = \alpha_{m1}f_1 + \alpha_{m2}f_2 + \dots + \alpha_{mk}f_k + e_m \end{cases}$$

式中， $x'_1 \sim x'_m$  是对原始变量进行均值为 0、标准差为 1 标准化后的变量； $f_i$  为第  $i$  个因子； $\alpha_{ij}$  为  $x'_i$  在共性因子  $f_i$  上的载荷，它的统计意义就是第  $i$  个变量与第  $j$  个公共因子的相关系数，表示  $x_i$  依赖  $f_j$  的份量。载荷的 SPSS 输出是在成分矩阵中，旋转后的载荷在旋转成分矩阵中。

## ② 公因子方差。

因为  $x'_1 \leq x'_m$  是原始变量  $x_1 \leq x_m$  标准化后的变量, 因此每个变量的方差均为 1, 即  $\text{Variance}(x'_i) = 1$ , 记作  $\text{Va}(x'_i)$ , 有

$$\text{Va}(x'_i) = \alpha_{i1}^2 + \alpha_{i2}^2 + \alpha_{i3}^2 + \cdots + \alpha_{im}^2 + V(e_i) = 1$$

它由两部分组成:

- 一部分是几个公共因子共同引起的公因子方差, 也称共同度  $h_i^2$ , 即

$$h_i^2 = \alpha_{i1}^2 + \alpha_{i2}^2 + \alpha_{i3}^2 + \cdots + \alpha_{im}^2$$

- 另一部分是由特殊因子引起的特性方差  $V(e)$ 。公因子方差占总方差的百分比越大, 说明公因子的作用越大。因为每个变量的方差均为 1, 因此公因子方差数值就是所占的百分比数值, 故又称分因子方差比。

要根据因子载荷和共性方差的大小解释共性因子  $f_i$  的意义, 须计算公因子方差

$$\text{Vc}(x'_i) = \sum_{j=1}^m \alpha_{ij}^2$$

如果取前  $k$  个因子, 公因子方差为

$$\text{Vc}(x'_i) = \sum_{j=1}^k \alpha_{ij}^2$$

③ 因子得分。因子得分就是每个观测的共性因子的值。要计算因子得分必须写出共性因子表达式。而共性因子不是能直接观测得到的, 它是潜在的, 但是可以通过可观测的变量获得, 即可以把共性因子表达成可观测变量的线性组合形式, 通常用回归方法解决。这样就可以通过每个观测的各变量值, 计算该观测的因子得分。

④ 关于旋转。要结合专业知识解释共性因子具有的实际意义并不是很容易的事, 常常得不到满意的解释。数学可以证明, 满足模型要求的共性因子并不唯一。只要对初始共性因子进行旋转, 就可以获得一组新的共性因子。所谓旋转就是一种坐标变换。在旋转后的新坐标系中, 因子载荷将得到重新分配, 使公因子负荷系数向更大(向 1)或更小(向 0)方向变化, 因此有可能对潜在因子作专业性解释, 使对公因子的命名和解释变得更加容易。对初始因子进行旋转的方法很多, 通常分为两类:

- 一类能保证旋转后各共性因子仍然正交, 称正交旋转。如方差最大正交旋转, 就是使共性因子上的相对载荷平方的方差之和达到最大, 并保证原共性因子之间的正交性和共性方差总和不变。
- 另一类旋转后不能保证各共性因子之间的正交关系, 如斜交旋转。

因子分析的一个重要目的在于对原始变量进行分门别类的综合评价。如果因子分析结果保证了因子之间的正交性(不相关), 但对因子不易命名, 可以通过对因子模型的正交旋转, 保证变换后各因子仍正交, 这是比较理想的情况。如果经过正交变换后对公因子仍然不易解释, 也可以进行斜交旋转, 或许可以得到比较容易解释的结果。

### 3. 因子分析过程的功能

SPSS 使用【因子分析】过程进行因子分析, 见图 14-1。主成分分析是作为因子分析的一种(没有旋转的)方法出现的。可以通过对话框指定因子提取的方法, 以及控制因子提取进程的

参数;可以指定旋转方法;可以对参与因子分析的变量给出描述统计量,指定输出负荷矩阵的格式;还可以产生新变量,其值是因子得分,并将其保存在数据文件中。使用【因子分析】过程的命令语句和一系列子命令还允许:

- (1) 一个命令完成多种方法的分析,对一种因子提取结果进行多种旋转。
- (2) 指定在提取因子与旋转时进行迭代的收敛判据,控制因子提取及旋转的进程。
- (3) 指定产生单个的旋转因子散点图。
- (4) 具体指定保存多少个因子。
- (5) 把相关矩阵或因子负荷矩阵写到磁盘上,以便进一步分析。
- (6) 指定主轴因子法的对角线上的值。
- (7) 从存储设备读取相关矩阵或因子负荷矩阵,并进一步分析。

4. 因子分析对变量的要求与假设

(1) 在因子分析中研究的是包含原始变量绝大部分信息的综合变量,对原始变量不分因变量和自变量。因子分析要求参与分析的变量必须是等间隔测度的或是比率的数值型变量。分类变量不适合作因子分析。那些明显可以作皮尔逊相关系数计算的数据才适合进行因子分析。观测应该彼此独立。一般观测数应为变量数的 5 倍以上。

(2) 因子分析的前提。因子分析模型指定变量由公因子(由模型估计的因子)和特殊因子(与原始观测变量不交迭)确定。参数计算的前提是,假设所有特殊因子彼此不相关,而且与公因子也不相关。

14.1.2 因子分析过程

对于初学统计分析的读者,可以完全使用系统默认值进行最简单的因子分析。虽然可能得不到非常满意的结果,但通过初步分析可以对所研究的问题有一个初步的认识,对进一步的分析会有帮助。对于比较简单的问题,有时只使用系统默认值进行因子分析就可以得到比较满意的结果。

【例 1】 数据文件 data14-01 中的数据是美国洛杉矶标准大城市统计区中的 12 个人口调查区的 5 个经济学指标(变量)的数据。下面以对 12 个地区的 5 个经济指标的调查数据进行因子分析为例,说明因子分析过程。



图 14-3 【因子分析】主对话框

- 1) 定义变量及标签  
no(编号)、pop(总人口)、school(中等学校平均校龄)、employ(总雇员数)、services(专业服务项目数)、house(中等房价)。
- 2) 使用默认值进行因子分析
  - (1) 读取数据文件 data14-01。按【分析→降维→因子分析】顺序单击菜单项,打开【因子分析】主对话框,见图 14-3。
  - (2) 指定参与分析的变量。

- (3) 在源变量框中选择 pop、school、employ、services、house 5 个变量,移到右边的【变量】框中。
- (4) 单击【确定】按钮,运行因子分析过程。



(5) 输出结果见表 14-4~表 14-6。

表 14-4 所示为公因子提取前与公因子提取后的公因子方差比表。

表 14-4 公因子方差比表  
公因子方差

	初始	提取
总人口	1.000	.988
中等校平均校龄	1.000	.885
总雇员数	1.000	.979
专业服务项目数	1.000	.880
中等房价	1.000	.938

提取方法：主成份分析。

表中的初始值是在提取因子(或成分,系统默认的是主成分法)之前的各变量的公因子方差比。对主成分分析来说,该值是要被分析的矩阵(相关矩阵或协方差矩阵)的对角线元素;对因子分析来说,这些值是用其他变量作为预测变量时每个变量的载荷的平方和。由于分析的是相关阵,原始变量的公因子方差均为 1(如果分析的是协方差阵,此处为各变量的方差),5 个变量的公因子方差比之总和为 5。

表中的提取值是各变量的未旋转的公因子方差比。表中的公因子方差比都很高,它表明提取的成分能很好地描述这些变量。

表 14-5 总方差分解

表 14-6 主成分分析的因子载荷阵

解释的总方差

成份矩阵<sup>a</sup>

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	2.873	57.466	57.466	2.873	57.466	57.466
2	1.797	35.933	93.399	1.797	35.933	93.399
3	.215	4.297	97.696			
4	.100	1.999	99.695			
5	.015	.305	100.000			

提取方法：主成份分析。

	成份	
	1	2
总人口	.581	.806
中等校平均校龄	.767	-.545
总雇员数	.672	.726
专业服务项目数	.932	-.104
中等房价	.791	-.558

提取方法：主成份。

a. 已提取了 2 个成份。

表 14-5 所示为各成分的公因子方差表。其中,“成份”中列出的是各成分的序号。“初始特征值”是相关矩阵或协方差矩阵的特征值。这些值是用于确定哪些因子(或成分)应保留,共有 3 项:

- “合计”中列出的是各成分的特征值。第一成分特征值为 2.873,第二成分特征值为 1.797。本例只有前两个因子的特征值大于 1。
- “方差的%”列出的是各成分所解释的方差占总方差的百分比,也就是各因子特征值占特征值总和的百分比。
- “累积%”自上至下列出的是各因子方差占总方差百分比的累积百分比。前两个因子的特征值之和占总方差的 93.4%。即前两个因子解释原始 5 个变量的 93.4%的变异。

“提取平方和载入”(应为“提取的因子载荷平方和”)是未经旋转的因子载荷的平方和。它给出的是每个因子(或成分)的特征值解释的方差占总方差的百分比和累计百分比。从初始分析的统计量可以看出按照系统默认值给出的分析原则,提取原则是特征值大于 1,那么应该取前两个因子(就本次分析来说也可称主成分)。而前两个因子已经对大多数数据给出了充分的概括,可以看出前两个成分所解释的方差占总方差的 93.4%。因此,最后结果是确定提取两个主成分。所以使用这些成分相当大程度上减少了原始数据的复杂性,仅丢失 6.6%的信息。

表 14-6 所示为(仅对不做旋转的主成分分析而言)因子载荷阵。它显示了原始变量与各主成分之间的相关程度。根据其相关程度的大小,综合出各因子的含义。

可以看出,第一主成分与 3 个变量的相关较高,这 3 个变量是“专业服务项目数”、“中等校平均校龄”和“中等房价”。而第二主成分与总人口数和总雇员数的相关更高些。

由以上输出结果可以认为对因子的提取结果是比较理想的。但是要想对两个因子命名就感

到比较困难，每个因子与原始变量相关系数没有很明显的差别。因此为了对因子进行命名，可以进行旋转，使系数向 0 和 1 两极分化。这就要使用选项了。

因子分析过程的各选项如下。

(1) 在主对话框中，单击【描述】按钮，打开如图 14-4 所示的对话框，从中选择描述统计量。

① 【统计量】栏。

- 【单变量描述性】。输出参与分析的原始变量的均值、标准差等单变量描述统计量。
- 【原始分析结果】(应为初始解)。这是系统默认选项。它给出因子提取前，分析变量的公因子方差。对主成分分析来说，这些值是分析变量的相关或协方差矩阵的对角线元素；对因子分析来说，是每个变量用其他变量作预测因子的载荷平方和。

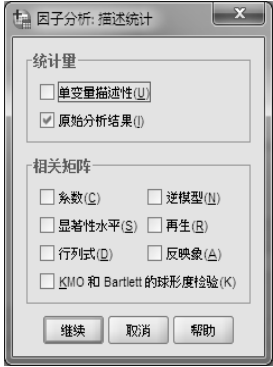


图 14-4 【因子分析: 描述统计】对话框

② 【相关矩阵】栏。

- 【系数】。原始分析变量间的相关系数矩阵。
- 【显著性水平】。每个相关系数等于 0 的单尾假设检验的显著性水平。
- 【行列式】。相关系数矩阵的行列式。
- 【逆模型】(应为逆矩阵)。相关系数矩阵的逆矩阵。
- 【再生】相关矩阵。此项给出因子分析后的相关矩阵，还给出残差，即原始相关与再生相关之间的差值。
- 【反映象】。给出反映象相关矩阵，包括偏相关系数的负数；反映象协方差矩阵，包括偏协方差的负数。在一个好的因子模型中除对角线上的系数较大外，远离对角线上的元素的系数应该比较小。
- 【KMO 和 Bartlett 的球形度检验】。要求进行 KMO 检验和球形 Bartlett 检验。选择此项给出对采样充足度的 Kaisex-Meyer-Olkin 测度，检验变量间的偏相关是否很小。Bartlett 球形检验，将检验相关矩阵是否单位矩阵，它表明因子模型是否不合适。也就是说，数据是否适合作因子分析。

(2) 在主对话框中，单击【抽取】(提取)按钮，打开如图 14-5 所示的对话框。

① 因子提取方法选项。

【方法】下拉列表给出一组提取方法的选项。提供 7 种提取方法供选择：

- 【主成分】法。该方法假设变量是因子的纯线性组合。第一成分有最大的方差，后续的成分可解释的方差逐个递减。主成分法是常用的获取初始因子分析结果的方法。它假设特殊因子作用可以忽略不计。
- 【未加权的最小平方方法】。该方法使观测的和再生相关矩阵之差的平方和最小，不计对角线元素。
- 【综合】(应为【广义】)【最小平方方法】。用变量值的倒数加权，使观测的和再生的相关矩阵之差的平方和最小。给较高值的权重比给较低值的权重要小。
- 【最大似然】法。此方法不要求多元正态分布。该方法给出参数估计。如果样本来自多元正态总体，它们与原始变量的相关矩阵极为相似。用变量单值倒数对原始分析变量加权。

● **【主轴因子分解】**。使用多元相关的平方作为对公因子方差的初始估计。初始估计公因子方差时，多元相关系数的平方置于对角线上。这些因子载荷用于估计新公因子方差，替换对角线上的前一次的公因子方差估计。每次迭代结束都计算从上次到本次迭代结果公因子方差的变化量。这样的迭代持续到公因子方差的变化量满足提取因子的收敛判据时为止。

● **【 $\alpha$  因子分解】**。 $\alpha$  因子提取法。

● **【映像因子分解】**。映像因子提取法(由 Guttman 提出)。根据映像学原理提取公因子，并把一个变量看作其他各变量的多元回归，而不是假设因子的函数。

② 在**【分析】**栏中指定分析矩阵。

● **【相关性矩阵】**(应为相关矩阵)，使用变量的相关矩阵进行提取因子的分析。如果参与分析的变量的测度单位不同时，应该选择此项。

● **【协方差矩阵】**。使用变量的协方差矩阵进行提取因子的分析。如果参与分析的变量测度单位相同，可以选择此项。

③ 在**【抽取】**栏中选择提取结果。理论上，因子数目与原始变量数目相等，但因子分析的目的是用少量因子代替多个原始变量。选择提取多少个因子由本组选项决定。

● **【基于特征值】**。指定提取的因子应该具有的特征值范围，在此项后面的矩形框中给出，系统默认值为 1，即要求提取那些特征值大于 1 的因子。指定特征值来决定提取因子数目的方法是系统默认的方法。

● **【因子的固定数量】**。指定提取公因子的数目。选择此项后，将指定的数目输入到该选项的矩形框中，数值应该是 0 至分析变量数目之间的正整数。

④ 在**【输出】**栏中指定与因子提取有关的输出项。

● **【未旋转的因子解】**。要求显示未经旋转的因子提取结果。此为系统的默认选项。

● **【碎石图】**。要求显示按特征值大小排列因子，以特征值为两个坐标轴绘制碎石图。该图有助于确定保留多少个因子。典型的碎石图有一个明显的拐点，在该点之前是与大因子有关的陡峭的折线，之后是与小因子有关的缓坡折线。

⑤ **【最大收敛性迭代次数】**框。指定因子分析停止的最大迭代次数，系统的默认值为 25，可以修改该值。

(3) 在主对话框中，单击**【旋转】**按钮，打开如图 14-6 所示的对话框。

① 在**【方法】**栏中选择旋转方法。

● **【无】**。不进行旋转。是系统默认的选项。

● **【最大方差法】**。方差最大旋转，是一种正交旋转。它使每个因子上的具有最高载荷的变



图 14-5 【因子分析：抽取】子对话框



图 14-6 【因子分析：旋转】对话框

量数最小, 因此可以简化对因子的解释。

- 【直接 Oblimin 方法】。直接斜交旋转, 指定此项可以在下面的矩形框中输入  $\delta$  值, 0 值产生最高相关因子。 $\delta$  值越接近 0, 斜交程度越深。大负数产生旋转的结果与正交接近。要不想  $\delta$  值为 0, 可以输入一个小于等于 0.8 的值。
- 【最大四次方法】。四次最大正交旋转。该旋转方法使每个变量中需要解释的因子数最少, 可以简化对变量的解释。
- 【最大平衡值法】。平均正交旋转, 是简化对因子解释的方差最大旋转方法与简化对变量解释的四次最大正交旋转方法的结合, 可以使在一个因子上有高载荷的变量数和变量中需要解释的因子数最少。
- 【Promax】。斜交旋转方法, 允许因子彼此相关。它比直接斜交旋转更快, 因此适用于大数据集的因子分析。

② 在【输出】栏中选择有关输出的选项。

- 【旋转解】。指定旋转方法后才能指定此项。将对正交旋转显示旋转后的因子矩阵、因子转换矩阵, 对斜交旋转显示旋转后的因子矩阵、因子结构矩阵和因子间的相关矩阵。
- 【载荷图】。因子载荷散点图。将给出以两两因子为坐标轴的各变量的载荷散点图。如果有两个因子, 给出各原始变量基于旋转成分矩阵表输出数据的散点图; 如果多于两个因子, 则给出前三个因子的三维因子载荷散点图; 如果只提取了一个因子, 则不会输出载荷散点图。

注意: 选择此项给出的是经旋转后的因子载荷图。

③【最大收敛性迭代次数】框。指定旋转收敛的最大迭代次数, 系统默认值为 25, 可以在框中输入指定值。

(4) 在对话框中, 单击【得分】按钮, 打开【因子得分】, 如图 14-7 所示的对话框。

①【保存为变量】。将因子得分作为新变量保存在数据文件中。每次分析产生一组新变量, 每次分析产生多少个因子, 就生成多少个新变量。新变量名的最后一个数字表示分析的序号。因子序号占倒数第三个字符的位置, 倒数第二个字符为“-”。在输出窗中给出对因子得分的命名和变量标签, 表明用以计算因子得分的方法。

② 在【方法】栏中指定计算因子得分的方法。选择【保存为变量】项, 激活【方法】栏中各项。

- 【回归】法。其因子得分的均值为 0, 方差等于估计因子得分与实际因子得分之间多元相关的平方。
- 【Bartlett】。巴特利特法。因子得分均值为 0。超出变量范围的特殊因子平方和被最小化。
- 【Anderson-Rubin】。安德森-鲁宾法。是为了保证因子的正交性而对巴特利特因子得分的调整。其因子得分的均值为 0, 标准差为 1, 且彼此不相关。

③【显示因子得分系数矩阵】。是标准化的得分系数。原始变量值进行标准化后, 可以根据该矩阵给出的系数计算各观测的因子得分。还显示协方差矩阵。

(5) 在主对话框中, 单击【选项】按钮, 打开如图 14-8 所示的对话框。

① 在【缺失值】栏中选择处理缺失值的方法。

- 【按列表排除个案】(删除样品记录)。在分析过程中对指定的分析变量中有缺失值的观测一律剔除。所有分析变量带有缺失值的观测都不参与分析。



图 14-7 【因子分析：因子得分】对话框

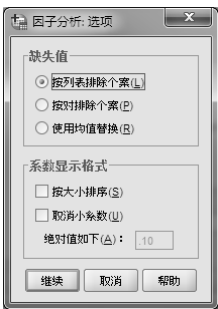


图 14-8 【因子分析：选项】对话框

- **【按对排除个案】**(成对删除)。成对剔除带有缺失值的观测，即在计算两个变量的相关系数时，只把这两个变量中带有缺失值的观测剔除。选择此项可以最大限度利用得来不易的原始数据。
- **【使用均值替换】**。用变量的均值代替该变量的所有缺失值。
- ② 在**【系数显示格式】**栏中决定载荷系数的显示方式。
- **【按大小排序】**。载荷系数按其数值的大小排列并构成矩阵，使在同一因子上具有较高载荷的变量排在一起，便于得出结论。
- **【取消小系数】**(应为不显示那些绝对值小于指定值的载荷系数)。在**【绝对值如下:】**框中输入 0~1 之间的数作为临界值，系统默认值为 0.10。选择此项可以突出载荷较大的变量，便于得出结论。

14.1.3 因子分析实例

**【例 2】** 仍使用**【例 1】**的数据进行因子分析。

1) 操作步骤

(1) 读取数据文件 data14-01。按**【分析→降维→因子分析】**顺序单击菜单项，打开**【因子分析】**对话框。

(2) 将 pop、school、employ、services、house 这 5 个变量移到**【变量】**框中。

(3) 在主对话框中，单击**【描述】**按钮，打开相应的对话框。

① 在**【统计量】**栏中选择要求输出的统计量

- 选择**【单变量描述性】**，要求显示单变量的描述统计量。
- 选择**【原始分析结果】**，要求显示初始因子分析结果。
- ② 在**【相关性矩阵】**栏中选择要求输出的相关矩阵。
- 选择**【系数】**，要求显示相关矩阵的相关系数。
- 选择**【显著性水平】**，要求显示针对相关系数为 0 的假设检验显著性概率。

(4) 在主对话框中，单击**【抽取】**按钮，打开相应对话框。

- ① 在**【方法】**后下拉列表中，选择**【主成分】**选项。
- ② 在**【分析】**栏中，选择**【相关性矩阵】**选项。
- ③ 在**【抽取】**栏中，选择**【因子的固定数量】**，并在其后矩形框中输入提取因子数“2”。
- ④ 在**【输出】**栏中选择要求的输出项：
  - 选择**【未旋转的因子解】**，在输出窗口中显示旋转前的因子分析结果。
  - 选择**【碎石图】**，在图表窗口中显示因子碎石图。

- ⑤ 在【最大收敛性迭代次数】框中选择停止迭代的最大迭代次数。使用默认值 25。
- (5) 在主对话框中单击【旋转】按钮，打开【旋转】对话框。
- ① 在【方法】栏中，选择【最大方差法】。
- ② 在【输出】栏中选择【旋转解】和【载荷图】，前者要求显示旋转后的结果，后者要求显示因子载荷图。
- (6) 在主对话框中，单击【得分】按钮，打开相应对话框。
- ① 选择【保存为变量】，以变量形式将因子得分保存在数据文件中，使用【方法】栏中默认的【回归】。
- ② 选择【显示因子得分系数矩阵】。
- (7) 在主对话框中，单击【选项】按钮，打开相应对话框。
- ① 在【缺失值】栏中选择【按列表排除个案】(删除样品)。
- ② 在【系数显示格式】栏中选择【按大小排序】。
- (8) 在主对话框中，单击【确定】按钮，执行运算。

2) 执行结果(见表 14-7~表 14-13，图 14-9~图 14-11)  
此处略去公因子方差表和各因子方差分解表，分别与表 14-4 和表 14-5 相同。

3) 结果解释、分析与结论

表 14-7 所示为单变量描述统计量。自左至右显示了变量标签、各变量的均值、各变量的标准差、参与计算这些统计量的观测数。

表 14-8 所示为各分析变量的相关矩阵。

表 14-7 单变量描述统计量

描述统计量			
	均值	标准差	分析 N
总人口	6241.67	3439.994	12
中等校平均校龄	11.442	1.7865	12
总雇员数	2333.33	1241.212	12
专业服务项目数	120.83	114.928	12
中等房价	17000.00	6367.531	12

表 14-8 原始变量的相关矩阵

		相关矩阵				
		总人口	中等校平均校龄	总雇员数	专业服务项目数	中等房价
相关	总人口	1.000	.010	.972	.439	.022
	中等校平均校龄	.010	1.000	.154	.691	.863
	总雇员数	.972	.154	1.000	.515	.122
	专业服务项目数	.439	.691	.515	1.000	.778
	中等房价	.022	.863	.122	.778	1.000
Sig. (单侧)	总人口		.488	.000	.077	.472
	中等校平均校龄	.488		.316	.006	.000
	总雇员数	.000	.316		.043	.353
	专业服务项目数	.077	.006	.043		.001
	中等房价	.472	.000	.353	.001	

图 14-9 所示为表现各成分特征值的碎石图。分析碎石图可以看出，因子 1 与因子 2，以及因子 2 与因子 3 之间的特征值之差值比较大。而因子 3、4、5 之间的特征值差值都比较小。可以初步得出结论：保留两个因子将能概括绝大部分信息。明显的拐点为 3，因此提取两个因子比较合适。证实了表 14-5 中的结果。

表 14-9 所示是初始提取的因子载荷矩阵。相关系数比较接近，不好命名。

表 14-10 所示为因子旋转的转换矩阵。

表 14-11 所示是旋转后的因子载荷矩阵。表下方是有关因子提取与旋转方法的说明：使用主成分法提取因子，使用最大方差法旋转，经 3 次迭代收敛。

表中给出了旋转后的因子(或成分)与原始变量的相关矩阵，是按系数由大到小排列的。可以看出，经过旋转后相关系数已经明显地发生变化了。第一个主成分对“中等房价”、“中等校平均校龄”、“专业服务项目数”有绝对值较大的相关系数，第二个因子相关系数绝对值较大的正好是 5 个原始变量中的另外 2 个，即“总人口和总雇员数”。根据这些变量的原始含义可以

对 2 个因子进行命名。第一个因子主要概括了一般的社会福利情况的因子，“中等房价”、“中等校平均校龄”和“社会服务项目数”可以命名为福利条件因子；第二个因子主要概括了人的情况，“总人口”和“总雇员数”，可以称为人口因子。

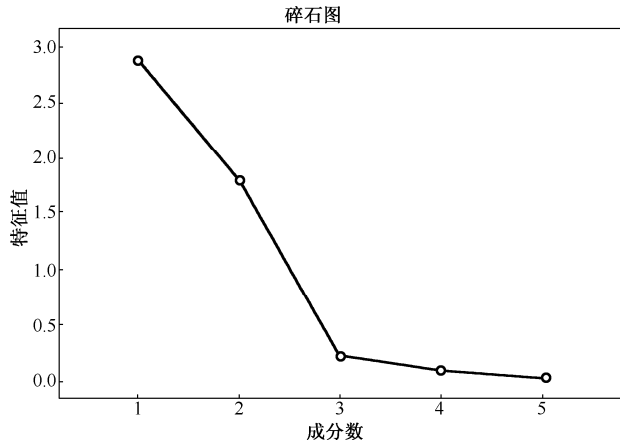


图 14-9 特征值碎石图

表 14-12 所示为因子得分系数矩阵。根据因子得分系数和原始变量的标准化值，可以计算每个观测的各因子的得分数，并可以据此对观测进行进一步的分析。旋转后的因子(主成分)表达式可以写成：

$$FAC1\_1 = -0.091 \times pop' + 0.392 \times school' - 0.039 \times employ' + 0.299 \times services' + 0.403 \times house'$$
$$FAC2\_1 = 0.484 \times pop' - 0.096 \times school' + 0.465 \times employ' + 0.138 \times services' - 0.098 \times house'$$

注意：因子表达式中的各变量均为经过均值为 0、标准差为 1 标准化后的变量。用原变量名加 “'” 表示。

表 14-9 旋转前因子载荷矩阵

	成份	
	1	2
总人口	.581	.806
中等校平均校龄	.767	-.545
总雇员数	.672	.726
专业服务项目数	.932	-.104
中等房价	.791	-.558

提取方法：主成份。  
已提取了 2 个成份。

表 14-10 转换矩阵

成份转换矩阵		
成份	1	2
1	.821	.571
2	-.571	.821

提取方法：主成份。  
旋转法：具有 Kaiser 标准化的正交旋转法。

表 14-11 旋转后因子载荷矩阵

	成份	
	1	2
总人口	.016	.994
中等校平均校龄	.941	-.009
总雇员数	.137	.980
专业服务项目数	.825	.447
中等房价	.968	-.006

提取方法：主成份。  
旋转法：具有 Kaiser 标准化的正交旋转法。

a. 旋转在 3 次迭代后收敛。

表 14-13 所示是估计回归因子分数的协方差矩阵，即因子(两个主成分)间的相关矩阵。可以看出旋转后第一主成分与第二主成分是完全不相关的。这也是因为使用最大方差正交旋转后因子间仍然正交。

图 14-10 所示为旋转后的因子(成分)载荷图，分别以第一主成分和第二主成分为横、纵轴坐标，按表 14-12 中数据作图得到主成分图(图中的指示线是作者加的)。从图中可以看出旋转后各成分的变量更集中了。

表 14-12 因子得分系数矩阵

成份得分协方差矩阵		
成份	1	2
1	1.000	.000
2	.000	1.000

提取方法:主成份。  
旋转法:具有 Kaiser 标准化的  
正交旋转法。  
构成得分。

表 14-13 估计回归因子分数的协方差矩阵

成份得分系数矩阵		
	成份	
	1	2
总人口	-.091	.484
中等校平均校龄	.392	-.096
总雇员数	-.039	.465
专业服务项目数	.299	.138
中等房价	.403	-.098

提取方法:主成份。  
旋转法:具有 Kaiser 标准化的正交旋  
转法。  
构成得分。

图 14-11 所示是在数据编辑窗口中,以新变量的形式保存的因子得分信息。数据文件中因子分数变量的命名:FAC1\_1 是第一次分析的第一个回归因子分数,FAC2\_1 标签是第 1 次分析的第二个回归因子分数变量。可以将此带有新变量的数据窗口中的数据保存为另一个数据文件 data14-01a。

有了观测的因子得分变量的值,可以进一步对观测估计因子得分变量进行聚类分析,进一步对每个调查区进行人口与福利方面的分类或分析。

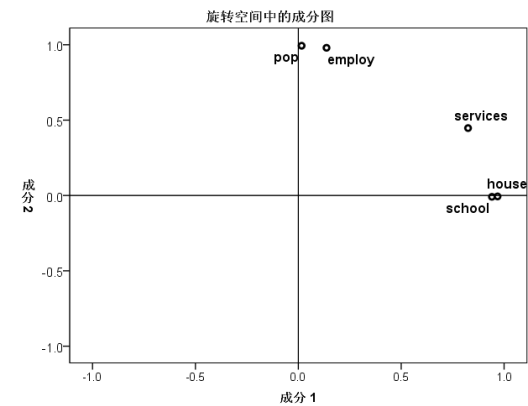


图 14-10 旋转后的因子载荷图



图 14-11 各观测的两个因子得分的新变量

14.1.4 利用因子得分进行聚类

【例 3】 本例是利用新变量对 12 个调查区进行聚类分析的过程及结果。聚类要求聚为 2 类、3 类、4 类,然后利用【图形】菜单功能作散点图,比较分为 2 类和 3 类的结果。

1) 操作步骤

(1) 完成因子分析后,在数据文件中保存有各观测的因子得分,见图 14-11,或直接读取数据文件 data14-01a。该文件保存有各观测因子得分。

(2) 按【分析→分类(聚类)→系统聚类】顺序单击菜单项,打开【系统聚类分析】主对话框。

(3) 在主对话框中:

- ① 指定 FAC1\_1 和 FAC2\_1 变量为分析变量,进入【变量】框中。
- ② 指定编号变量作为标识变量,进入【标注个案】框中。
- ③ 在【聚类】栏中选择【个案】,要求进行样品聚类。
- ④ 【输出】栏中选择【统计量】和【图】。



(4) 在主对话框中，单击【统计量】按钮，打开【系统聚类分析：统计量】对话框。选中【相似性矩阵】。在【聚类成员】栏中选择【方案范围】，并在【最小聚类数】框中输入“2”，在【最大聚类数】框中输入“4”，即要求聚为 2 类到 4 类的结果。

(5) 在主对话框中，单击【绘制】按钮，打开【系统聚类分析：图】对话框。

① 选中【树状图】。

② 在【冰柱】栏中指定要在冰柱图中出现的类的范围。选中【聚类的指定全距】来指定类范围，在【开始聚类】框中输入起始类数“2”，在【停止聚类】框中输入终止类数“4”，在【排序标准】(应为【步长】)框中输入步长“1”。

③ 在【方向】栏中选择【垂直】。要求显示方向是纵向的。

(6) 在主对话框中，单击【方法】按钮，打开【系统聚类分析：方法】对话框。

① 在【聚类方法】下拉列表中选择【组间联接】。

② 在【度量标准】栏中的【区间】(应为等间隔测度)下拉列表中选择【平方 Euclidean 距离】，要求根据两个因子间的欧氏距离的平方进行聚类。

③ 在【转换值】栏中的【标准化】下拉列表中选择默认值【无】，因为两个因子得分本身就是根据标准化变量得出的无量纲变量。

(7) 在主对话框中单击【保存】按钮，打开【系统聚类分析：保存】对话框。在【聚类成员】栏中选择【方案范围】，并在【最小聚类数】框中输入“2”；在【最大聚类数】框中输入“4”，即要求保存 3 个新变量，表示聚为 2 类、3 类、4 类时每个观测各归为哪一类。

通过作散点图观察 12 个调查区的经济情况分析。对各选项的含义请参考第 13 章的有关内容。

(8) 在主对话框中，单击【确定】按钮，执行运算。

2) 输出结果(见表 14-14、表 14-15 和图 14-12、图 14-13)

非关键的输出表没有列出。

表 14-14 相似性矩阵

案例	平方 Euclidean 距离											
	1:1	2:2	3:3	4:4	5:5	6:6	7:7	8:8	9:9	10:10	11:11	12:12
1:1	.000	5.297	6.606	.595	.607	6.231	4.186	3.289	1.670	1.255	5.605	4.036
2:2	5.297	.000	.740	3.491	2.933	4.053	.269	4.533	6.234	10.801	5.583	6.249
3:3	6.606	.740	.000	5.638	4.825	1.733	1.802	2.925	5.319	11.256	3.088	4.168
4:4	.595	3.491	5.638	.000	.033	7.286	2.212	4.767	3.556	3.534	7.342	6.084
5:5	.607	2.933	4.825	.033	.000	6.396	1.840	4.143	3.190	3.608	6.523	5.431
6:6	6.231	4.053	1.733	7.286	6.396	.000	5.518	.649	2.521	8.075	.249	.910
7:7	4.186	.269	1.802	2.212	1.840	5.518	.000	5.290	6.302	9.673	6.911	7.141
8:8	3.289	4.533	2.925	4.767	4.143	.649	5.290	.000	.612	4.155	.307	.139
9:9	1.670	6.234	5.319	3.556	3.190	2.521	6.302	.612	.000	1.590	1.607	.635
10:10	1.255	10.801	11.256	3.534	3.608	8.075	9.673	4.155	1.590	.000	6.367	4.121
11:11	5.605	5.583	3.088	7.342	6.523	.249	6.911	.307	1.607	6.367	.000	.265
12:12	4.036	6.249	4.168	6.084	5.431	.910	7.141	.139	.635	4.121	.265	.000

这是一个不相似矩阵

从输出信息很难看出各调查区在经济特性方面的区别。5 个变量转变为 2 个综合指标，即 2 个因子的好处在于减少了指标数目(降维)，而综合指标包含的信息没有损失多少。使用 2 个综合指标可以对调查区的经济状况更清楚地进行分析，还可以使用其他 SPSS 过程进行进一步的分析。

3) 利用因子得分变量作散点图

(1) 按【图形→旧对话框→散点/点状】顺序单击菜单项，打开【散点/点状】对话框，选

表 14-15 聚为 2 类、3 类、4 类的结果

群集成员			
案例	4 群集	3 群集	2 群集
1: 1	1	1	1
2: 2	2	2	2
3: 3	2	2	2
4: 4	1	1	1
5: 5	1	1	1
6: 6	3	3	1
7: 7	2	2	2
8: 8	3	3	1
9: 9	3	3	1
10: 10	4	1	1
11: 11	3	3	1
12: 12	3	3	1

择【简单分布】项，单击【定义】按钮后，打开【简单散点图】对话框，见图 14-14。

(2) 选择因子 2 的得分 FAC2\_1 作为 Y 轴变量送入【Y 轴】框中；选择因子 1 的得分 FAC1\_1 作为 X 轴变量送入【X 轴】框中；每个观测用它们的编号标识，故将变量 no 送入【标注个案】框中；每个观测按所属类别，使用不同的颜色或符号区分，先作聚为两类的散点图，将变量 Clu2\_1 送入【设置标记】框中。按 Clu2\_1 (标签为 Average Linkage Between groups) 的类数确定符号的类数。

(3) 单击【选项】按钮，打开相应对话框，见图 14-15。选择【使用个案标签显示图表】。

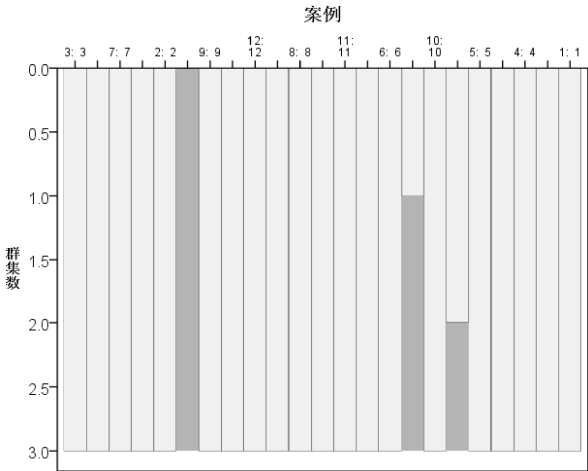


图 14-12 平均连接法形成的冰柱图

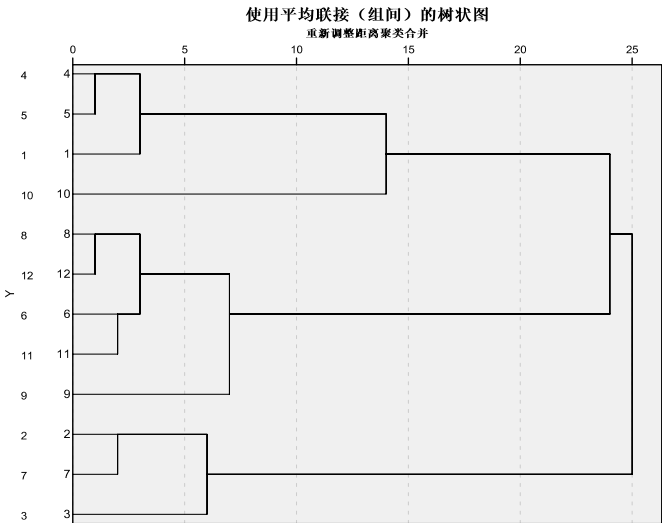


图 14-13 反映聚类全过程的树形图



图 14-14 【简单散点图】对话框

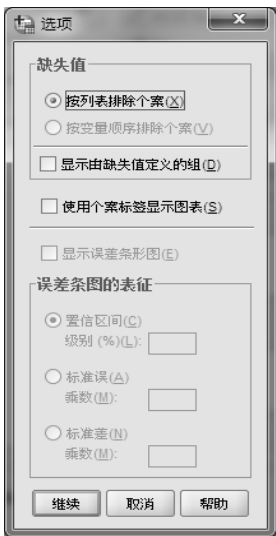


图 14-15 【选项】对话框

注意：选择此项，【标注个案】指定的变量值会标在散点图中观测点旁。

(4) 在主对话框中单击【确定】按钮，在输出窗中生成的散点图如图 14-16 所示。

(5) 再把变量 Clu3\_1 移入【设置标记】栏，代替 Clu2\_1，得到图 14-17。

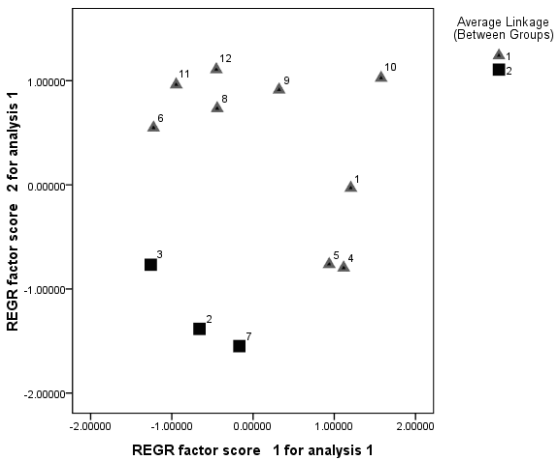


图 14-16 聚为 2 类的因子得分散点图

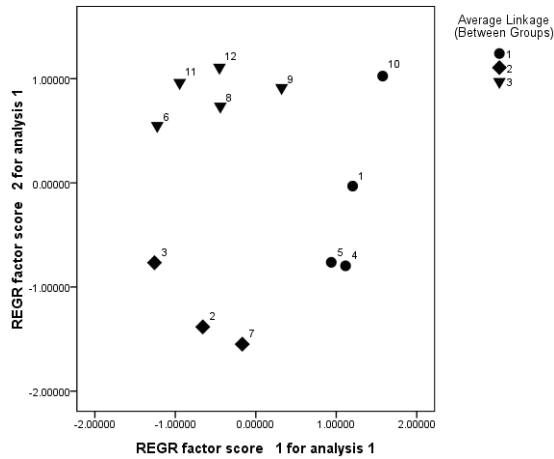


图 14-17 聚为 3 类的因子得分散点图

注意：图 14-16 和图 14-17 都是经过编辑的图形，为了弥补黑色印刷符号以深浅代替不同颜色而不易观察的缺点，改变了其中的分类标识符。读者在彩色显示器上可以分辨不同颜色的统一符号标识的分类，不用再编辑。

从图 14-16 可以看出，如果将调查区分为两类，第 2、3、7 区类号为 2 的，是福利因素和人口因素均比较低的；其余调查区的这两个因素水平比较高，可以认为经济状况是相对来说比较好的。

从图 14-17 可以更细致地划分和分析各调查区的经济水平。

① 类号为 2 的调查区有编号为 2、3、7 的三个地区，在图的左下角是两个因子得分均比较低，可以认为从 5 个经济指标来看均较差的地区。

② 类号为 3 的调查区 FAC1\_1 比较低，即福利因子得分较低；而 FAC2\_1 比较高，即人口

因子得分较高,说明总人口多,就业人数多,但反映福利的“中等校平均校龄”、“服务项目数”、“中等房价”均比较低。这样的地区有 6、8、9、11、12 号地区。

③ 类号为 1 的调查区位于散点图的右偏上方,可以看作人口和就业人数均较少、福利条件比较好的地区,有编号为 1、4、5 号地区。

④ 如果分为 4 类,则右上角的点将单独分为一类,是两个因子得分均较高的地区。用户可以自己根据因子得分聚 4 类并作散点图。

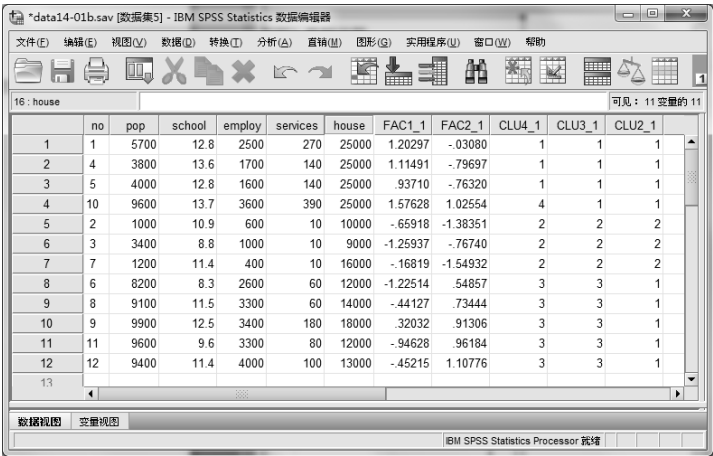


图 14-18 排序后的数据

通过以上分析可以看出,使用因子得分绘制出散点图是比较容易进行的,反过来对照原始数据,也可以得出同样的分析结论。但是直接使用 5 个原始变量来分类,就不够直观。由此对因子分析的作用也可以有较实际的体会了。

4) 排序后观察因子和原始数据

(1) 按【数据→排序个案】操作,将 Clu3\_1 作为排序关键字,按【升序】排序。排序后的数据保存在数据文件 data14\_01b 中。

(2) 观察各类原始变量的特点也可以对各类特征得出明显的结论,见图 14-18。

应该说明的是,实际应用中,作因子分析要求观测数至少应该是变量数的 5 倍以上。而本例 5 个变量仅 12 个观测,所以仅作为一个介绍方法的例题而已。

14.1.5 市场研究中的顾客偏好分析

在市场研究中,常常要求分析顾客的偏好和当前市场的产品与顾客偏好之间的差别,从而找出新产品开发的方向。顾客偏好分析时常用到主成分分析方法。

下面例题数据文件 data14-02 来自 SAS 公司。1980 年一个汽车制造商在竞争对手中选择了 17 种车型,访问了 25 个顾客,要求他们根据自己的偏好对 17 种车型打分,打分范围为 0~9.9,9.9 表示最高程度的偏好。

1) 数据文件格式

数据文件是以 25 个顾客的评分分为 25 个变量,即  $v_1 \sim v_{25}$ ,每种车型的 25 个分数是一个观测,17 种车型为 17 个观测。

2) 操作步骤

(1) 按【分析→降维→因子分析】顺序单击菜单项,打开【因子分析】主对话框。

- (2) 选择  $v_1 \sim v_{25}$  为分析变量送到【变量】栏中。
- (3) 在主对话框中单击【抽取】按钮，在相应的对话框中：
  - ① 在【方法】下拉列表中选择【主成分】分析方法。
  - ② 在【分析】栏中选择【相关性矩阵】，分析相关矩阵。
  - ③ 在【抽取】栏中选择【因子的固定数量】，并输入“3”。
  - ④ 在【输出】栏中选择【未旋转的因子解】，显示未旋转的因子结果；同时选择【碎石图】，要求作出特征值的碎石图。
  - ⑤ 【最大收敛性迭代次数】框使用默认值 25，结束迭代的判据为到达最大迭代次数 25。

(4) 在主对话框中单击【得分】按钮。在相应的对话框中选择【保存为变量】，并在【方法】栏中选择【回归】，要求通过回归方法计算因子得分并把因子得分作为变量保存到数据文件中。

(5) 单击【描述】按钮，在对话框中的【统计量】栏内不选择【原始分析结果】。

- (6) 在主对话框中单击【确定】提交系统执行。
- 3) 输出结果(见表 14-16、表 14-17、图 14-19、图 14-20)
- 4) 结果说明

由于选择了提取公因子的方法为主成分法，因此输出中的因子即成分。

表 14-16 初始因子载荷阵

	成份		
	1	2	3
被访者1	.274	.625	.330
被访者2	.956	.068	-.210
被访者3	.778	-.300	-.151
被访者4	.491	.735	.343
被访者5	.451	.698	-.318
被访者6	.238	.677	-.059
被访者7	.783	-.212	.170
被访者8	.510	-.051	.713
被访者9	-.513	.718	-.189
被访者10	.936	-.191	.050
被访者11	.852	.143	-.260
被访者12	.836	-.085	-.356
被访者13	.943	.000	-.149
被访者14	.830	.198	-.081
被访者15	.858	-.174	-.067
被访者16	-.015	.803	.077
被访者17	.105	.658	.235
被访者18	.717	.609	.096
被访者19	.779	.126	-.033
被访者20	.773	-.570	.124
被访者21	.071	.657	-.095
被访者22	.238	-.459	.753
被访者23	-.766	.333	.281
被访者24	-.162	-.753	-.209
被访者25	-.765	.158	-.270

提取方法：主成份。  
已提取了 3 个成份。

表 14-17 前 3 个因子(或成分)的方差解释

成份	提取平方和载入		
	合计	方差的 %	累积 %
1	10.837	43.348	43.348
2	5.802	23.207	66.555
3	2.060	8.240	74.795

提取方法：主成份分析。

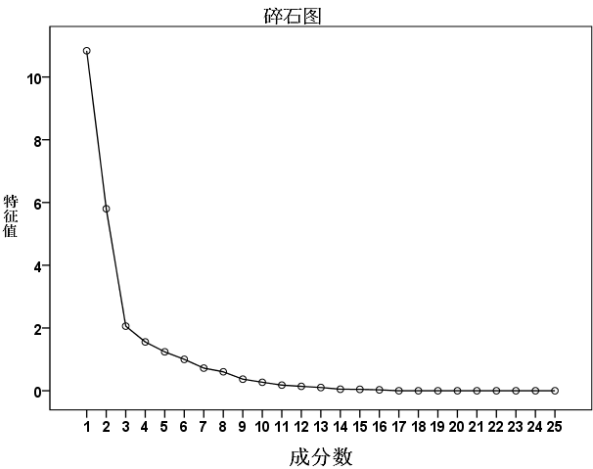


图 14-19 特征值碎石图

表 14-16 所示是提取的 3 个因子的因子载荷矩阵。行列交叉点上的数据是对应因子在变量(顾客)上的载荷。它体现了交叉点对应的因子(列)与对应变量(行)的相关程度。

在选择提取公因子的数量时，没有选择特征值大于 1 决定公因子数的方法，而是选择了提

取前 3 个公因子。因此表 14-17 为前 3 个因子所解释的原始变量的总方差，及其占总方差的百分比和累计百分比。可以看出，前 3 个因子(或成分)可以解释总方差的近 75%，其余 22 个因子只占 25%，可以说 3 个因子可以解释总方差的绝大部分。

图 14-19 所示是特征值碎石图。可以看出，前 3 个特征值间的差异很大，其余的变化很小，虽然也有特征值大于 1 的，但变化量很小，可见，取前 3 个因子是正确的。

图 14-20 所示是当前数据文件。其中最右边的变量 Fac1\_1、Fac2\_1 和 Fac3\_1 是各观测(17 种车型)的因子得分变量。该数据保存到数据文件 data14-02a 中。

data14-02a.sav [数据集合] - IBM SPSS Statistics 数据编辑器

文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 窗口(W) 帮助(H)

打印(P) 撤消(X) 重做(O) 剪切(C) 复制(C) 粘贴(P) 删除(D) 查找(F) 替换(R) 拼写(S) 窗口(W) 帮助(H)

数据编辑器 数据视图 数据源

图 14-20 数据文件中的 3 个新变量：因子得分

根据数据文件中的因子得分变量和表 14-16 中的数据作散点图，得到偏好图。  
5) 作偏好图

根据数据文件中的前两个因子得分变量作 17 个车型的散点图。也可先读取数据文件 data14-02a，步骤如下：

(1) 按【图形→旧对话框→散点/点状】顺序单击菜单项，打开【散点/点状】对话框。选择左上角的【简单分布】项，单击【定义】按钮，打开【简单散点图】对话框，见图 14-14。

(2) 将 Fac1\_1 (REGR factor scor-1 for analysis 1 [fact1-1]) 送入【X 轴】作为 X 轴变量，将 Fac2\_1 (REGR factor score 2 for analysis 1 [fact 2 1]) 送入【Y 轴】作为 Y 轴变量，变量 nametype 送入【设置标记】框中。单击【确定】按钮得到如图 14-21 所示的散点图。图中的字母均为变量 nametype 的值。

(3) 根据表 14-16 的因子载荷数据，见数据文件 data14-02b。用同样方法作另一个散点图，见图 14-22。

6) 分析与结论  
结合输出表，比较图 14-21 和图 14-22。如果有条件，可以将两张图的坐标原点对齐，并经过透明处理，则更便于比较。可以看出：

(1) 图 14-21 是根据 17 种车型的前两个因子得分作的图。  
第一因子反映了车的产地。分数最高的是(DL)沃尔沃，最低的是(P)福特。横坐标右端多为欧洲车(D、R)两种大众车或日本车(A、CI)两种本田车，左端多为美国车(P)福特、(CH)雪伏龙等，各自的第一因子得分说明顾客对欧洲车和日本车的评价较高。

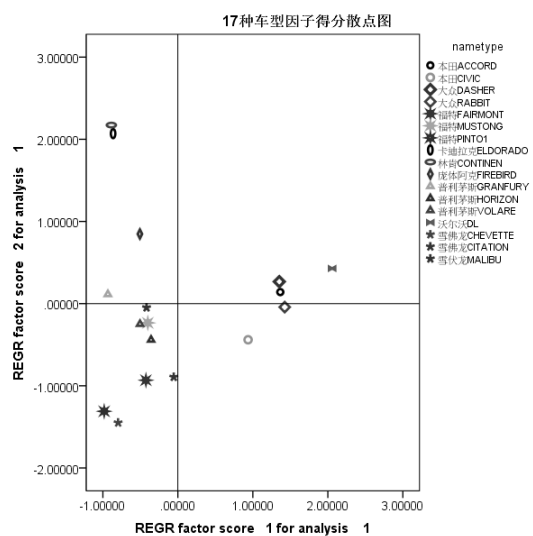


图 14-21 17 种车型的因子得分散点图

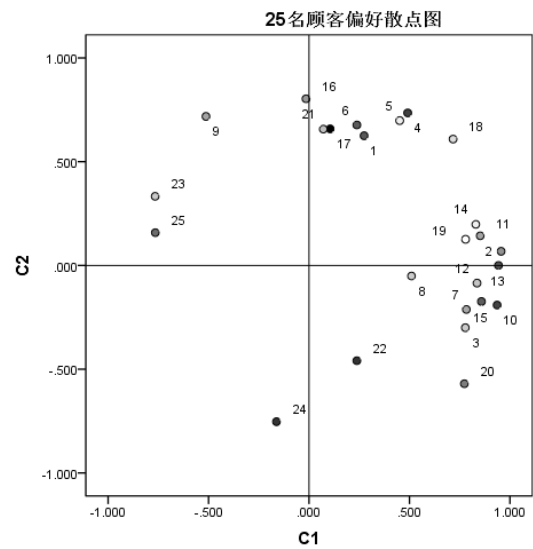


图 14-22 25 个顾客的偏好散点图

第二因子反映了车的特性：质量、动力、座位数等。分数高的是 (Co) 林肯、(E) 凯迪拉克，位于纵坐标高端，分数低的为 (P) 福特、(CH) 雪佛兰，说明顾客对高档车的质量评价较高。

(2) 与图 14-21 对应着进行综合分析，图 14-22 的点在第二象限的(左上方)的顾客偏好大型豪华美国车；点在第四象限的顾客偏好日本和欧洲车；第三象限的点很少，说明顾客中偏好美国小型车的很少。图 14-22 中，第一象限点很多，但相应图 14-21 中第一象限的车很少。这可能预示着新车型产品市场或该汽车生产商的主要竞争对手没有相应的产品。这正是新产品开发的方向：高质量、豪华大型欧洲、日本车。

1980 年的此项研究虽然使用的样本很少，但事后发现，这项研究对市场的分析相当准确：目前除美国福特外，其他中小型车 (MU 点) 几乎都不生产了；日本车在 20 世纪 80 年代的市场占有率是比较高的；日本和欧洲汽车制造商开发了大型豪华车，如德国的宝马、日本的凌志，都是很受顾客欢迎的车型。

## 14.2 对 应 分 析

### 14.2.1 对应分析概述

#### 1. 对应分析的思路

对应分析也称为相应分析，是在 R 型和 Q 型因子分析的基础上发展起来的一种多元统计方法。它首先由法国统计学家 J. P. Beozecri 于 1970 年提出。

因子分析根据研究对象的不同而分为研究指标(变量)的 R 型因子分析和研究样品的 Q 型因子分析，使用因子分析方法时这两个过程只能分开进行。这样，一方面会漏掉一些变量和样品间的信息，另一方面因子分析要求样品数必须是变量数的 5 倍，因此，在作 Q 型因子分析时，还要作比 R 型因子分析计算量更大的计算工作。因此从研究设计的要求来说，并不是最佳的，所以，有必要改良算法，从而达到总计算量最小而又同时考虑变量和样品的关系。

对应分析借助列联表独立性检验中卡方统计量的计算方法，对原始数据矩阵进行转换，公式是

$$p_{ij} = x_{ij} / \sum_i \sum_j x_{ij}$$

由此得到一个规格化的“概率”矩阵，使数据资料具有对称性，当数据资料具有对称性时，量纲的差异也被消除，R 型和 Q 型因子分析之间就建立起了联系，在作 R 型因子分析时也就同时完成了 Q 型因子分析的工作，克服了由于样品容量大带来的 Q 型因子分析计算量大的困难。

另外，根据 R 型因子分析和 Q 型因子分析的内在联系，可在同一个坐标轴图形中将变量和样品同时反映出来，图形中邻近的变量点表示它们关系密切，可分为一类，同样，邻近的样品点表示它们关系密切，可归为一类，而且属于同一类型的样品点可用邻近的变量点来表征。

对应分析的目的之一是在同时描述各个变量类别之间的关系时，在一个低维度空间中对对应表里的两个名义变量之间的关系进行描述。对每个变量而言，图中类别点之间的距离反映了邻近有相似分类图的各类别之间的关系。一个变量在从原点到另一个变量分类点的向量上的投影点描述了变量之间的关系。

很多学者认为对应分析方法是探索性数据分析的内容，因此，极大部分的使用者只要能够理解对应分析行、列记分图所包含的信息即可。

2. 对应分析中需要考虑的事项

- (1) 数据。用于分析的分类变量为名义测度变量。对合计数据或对除频数以外的对应测度，使用正相似性值的权重变量。
- (2) 有关程序。如果包含的变量超过两个，使用多重(多元)对应分析。如果是有序测度变量，则使用分类主成分分析。

3. 对应分析中的几个常见术语

- (1) 行(列)中心化。它是对行(列)变量中的原始数据进行转换的一种方法，用行(列)变量中的每个观测值减去行(列)观测值的均值来实现。
- (2) 行(列)边际(缘)。是指某一个行(列)变量中的观测值的总和。
- (3) 归一化处理。用行(列)变量中的每个观测值除以行(列)变量中的观测值总和进行数据转换的一种方法，由于转换后的行(列)变量中的观测值之和为 1 故得名。因此，使行(列)边际相等，实际上就是先对行(列)变量中的数据作归一化处理。
- (4) 质量。是指转换后数据的行与列的边缘概率。
- (5) 惯量。为每一维到其重心的加权距离的平方，用来度量行列关系的强度。
- (6) 奇异值。是惯量的平方根，反映了行与列各水平在二维图中分量的相关程度，是行与列进行因子分析产生新的综合变量的典型相关系数。
- (7) 惯量比例。是各维度分别解释总惯量的比例及累计百分比。
- (8) 对应表。行、列变量各类别组合在一起形成的各组合类别观测的分布表。

14.2.2 对应分析过程

1. 对应分析数据预处理

在对应分析中，原始数据必须整理成交叉表的单元格计数形式。在对应分析前先用【加权个



案】命令进行处理。因此，在对应分析的数据文件中，需定义 3 个变量。将要放在对应分析过程中的行和列的变量是分类变量。第 3 个变量是对应行、列的实际测试值，一般为尺度变量。

在进行对应分析前，应先用【数据】菜单中【加权个案】功能定义权重变量，方法参见第 2.4.1 节相关内容。如果权重变量中有 0 值，会发出警告，但不影响对应分析的正常分析工作。

## 2. 操作步骤

(1) 按【分析→降维→对应分析】顺序打开如图 14-23 所示的【对应分析】主对话框。

(2) 从变量表中选择行、列变量，分别送入【行】、【列】框中。

(3) 单击【定义范围】按钮，见图 14-24，可定义行(列)变量参与分析的分类范围。

① 【行变量的分类全距：C1】在【最小值】框中输入分类的最小值；在【最大值】框中输入最大值。这两个值必须是整数，否则在分析中会删除小数部分。单击【更新】按钮，可将定义的分类数据上传到【类别约束】框中。在分析中忽略在指定范围以外的分类值。

② 【类别约束】栏。定义类别的等同约束。所有类别最初没有约束。可以约束某个行(列)类别去等于其他的行(列)类别，或者定义一个行(列)类别作为辅助行(列)类别。如果分类值所代表的类别不符合分析需要或者界限是模糊的，可以使用等同约束将这样的类视为等同，即有相等记分的类。它共有 3 个选项：

- 【无】。默认选项，即分类数据保持原状，不作任何约束。



图 14-23 【对应分析】主对话框

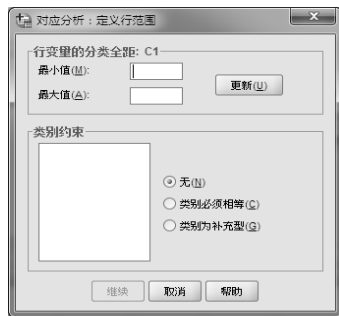


图 14-24 【对应分析：定义】对话框

- 【类别必须相等】。类别必须有相等的得分。假如类别的次序是不想要的或违反直觉的，则可以使用等同约束。这可从【更新】产生的分类值列表中选择类别，指定等同约束，至少有两个类别必须是相等的。能用等同约束的行(列)分类的最大数量是有效行(列)类别总数减 1。

- 【类别为补充型】。从【更新】产生的分类值列表中选择类别指定辅助类别。辅助类别不影响分析，只在由有效分类定义的空间里描述。辅助类别在定义的维度数里不扮演角色。最大辅助的行(列)类别的数量是行(列)类别总数减 2。

(4) 指定对应分析模型。单击【模型】按钮进入【对应分析：模型】对话框，见图 14-25。允许指定维度数、距离测度、标准化方法和常态化方法。

① 【解的维数】框。指定对应分析解的维度数，默认值为 2。通常选择使用较少的维度数来解释大多数的变差。最大维度数取决于用于分析的有效分类数和等同约束数。最大维度数是下列中较小的一个：

- 有效的行分类数减去被等同约束的行分类数加约束的行分类集数；

- 有效的列分类数减去被等同约束的列分类数加约束的列分类集数。

②【距离度量】栏。选择对应表的行间距离和列间的距离测度。

- 【卡方】。卡方距离测度用加权距离，这里的权重就是行或列的质量(边际概率)。标准对应分析要求使用卡方距离测度。本选项是系统默认方法。
- 【Euclidean】(欧氏距离)。用两行之间或两列之间的差的平方和的平方根作为距离测度。

③【标准化方法】栏。标准化方法选项，共 5 种。

- 【行和列均值已删除】(行和列两者变换中心化)。标准对应分析需用本方法。当选用【卡方】作为【距离度量】的选项时，系统只默认本方法。
- 【行均值已删除】。只有行作中心化变换。
- 【列均值已删除】。只有列作中心化变换。
- 【使行总和相等，删除均值】。先使行边际相等，再中心化行。
- 【使列总和相等，删除均值】。先使列边际相等，再中心化列。

④【正态化方法】(应为常规方法)栏。可从下列 5 个选项中选择一种：

- 【对称】法。对各个维度，行记分是列记分除以匹配奇异值的加权平均，列记分是行记分除以匹配奇异值的加权平均。使用本方法可以检查两个变量分类间的差异或相似性。
- 【主要】。行点和列点之间的距离是与选定的距离测度一致的对对应表中距离的近似值。如果要检查一个或两个变量的类别之间的差异，而不是两个变量之间的差异时，使用本方法。
- 【主要行】。行分数间的距离是在对应表中根据选定方法对距离测度的近似值。行记分是列记分的加权平均。要检查行变量的类间差异或类似程度时，使用本方法。
- 【主要列】。列分数间的距离是在对应表中根据选定方法计算的距离的近似值。列记分是行记分的加权平均。要检查列变量的类间差异或类似程度时，使用本方法。
- 【设定】。必须在 $\pm 1$ 间指定一个值。值“-1”对应于【主要列】，值“1”对应于【主要行】，值“0”对应于【对称法】。所有其他值传达行和列分数变化程度的惯量。本方法通常用来制作特制的二维图形。

(5) 单击【统计量】按钮，进入【对应分析：统计量】对话框，见图 14-26，指定输出哪些结果表。

- ①【对应表】。要求输出含有变量行和列边际总和的交叉分组列表。
- ②【行点概览】(行分数综述)。要求输出行综合表，表中包括行变量各分类的得分、质量、惯量、分数对维度惯量的贡献、维度对分数惯量的贡献。
- ③【列点概览】(列分数综述)选项。在输出窗中为各个列分类显示包括得分、质量、惯量、分数对维度惯量的贡献、维度对分数惯量的贡献的综合表。两个输出选项：



图 14-25 【对应分析：模型】对话框



图 14-26 【对应分析：统计量】对话框

④【对应表的排列】。输出按第一维度上得分的递增顺序排列的行、列对应表。在任意选项中，可为将要产生的序列改变的表指定最大维度数，为各维度产生一个从 1 到指定数目的序列改变表。

⑤【行轮廓表】。行归一化处理后的分布表。

⑥【列轮廓表】。列归一化处理后的分布表。

⑦【置信统计量】栏。在本选择中共有两个选项：

- 【行点】(行分数)。输出包括标准差和所有非辅助行分数相关内容的表格。
- 【列点】(列分数)。输出表格包括标准差和所有非辅助列分数相关内容。

(6) 统计图选项。单击【绘制】按钮，进入【对应分析：图】对话框：

图 14-27 【对应分析：图】对话框，见图 14-27。

①【散点图】栏。产生矩阵的所有维度的成双图。

- 【双标图】。双维图法。输出矩阵的行、列分数联合图。如果选择了【主要】这个常态化方法，则本选项无效。
- 【行点(行分数)】。输出矩阵的行分数的图。
- 【列点(列分数)】。输出矩阵的列分数的图。
- 【散点图的标识标签宽度】框。设置散点图中的 ID 标识宽度，默认值为 20。该值必须是小于等于 20 的正整数。

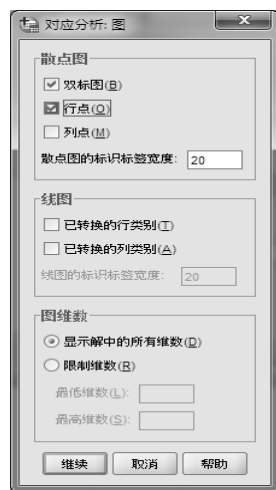


图 14-27 【对应分析：图】对话框

②【线图】栏。产生所选变量每一个维度的线图。有以下两个线图供选择。

- 【已转换的行类别】。输出行分类转换图。行分类值取决于相应的行记分。
- 【已转换的列类别】。输出列分类转换图。列分类值取决于相应列记分。
- 【散点图的标识标签宽度】框。指定线图 ID 标识宽度，默认值为 20。指定值必须是小于等于 20 的正整数。

③【图维数】栏。允许去控制在输出中显示的图的维度。

- 【显示解中的所有维数】。在散点图矩阵里显示解中的所有维度。
- 【限制维数】。显示的维度数受到成对图数的限制。如果限制维度，则必须选择作图的最低和最高维度。最低维度可从 1 到解的维度数减 1 的范围中取值，并且所作的图以较高维度为背景；最高维度可从 2 到解的维度数的范围中取值，并指出被使用于成对维度图中的最高维度。本说明适用于所有要求的多维图。

### 14.2.3 对应分析实例

【例 4】用对应分析的方法研究我国部分省份的农村居民人均消费支出结构。数据资料来源于《中国统计年鉴(1997 年)》。

在数据文件 data14-03 中，共有 3 个变量，分别为 province(省份：1 山西、2 内蒙古、3 辽宁、4 吉林、5 黑龙江、6 海南、7 四川、8 贵州、9 甘肃、10 青海)，consumption(消费支出分类：1 食品、2 衣着、3 居住、4 家庭设备及服务、5 医疗保健、6 交通和通讯、7 文教娱乐)，proportion(各种消费支出比例)是尺度变量。前两个为名义变量。

首先用【数据→加权个案】过程中的【加权个案】选项将 proportion(各种消费支出比例)设置为频数变量。

- (1) 按【分析→降维→对应分析】顺序单击菜单项，进入【对应分析】主对话框。
- (2) 选择 province(省份)，移入【行】框中，单击其下的【定义范围】按钮，在【最小值】框中输入“1”，在【最大值】框中输入“10”，单击【更新】按钮，送到【类别约束】框中。由于没有辅助项、等同约束项及强制性等同约束项，因此，在【类别约束】栏中使用系统默认选项【无】。单击【继续】按钮，返回主对话框。
- (3) 选择 consumption，移入【列】框中，单击【定义范围】按钮，在【最小值】框中输入“1”，在【最大值】框中输入“7”，单击【更新】按钮，送到【类别约束】框中。单击【继续】按钮返回【对应分析】主对话框。
- (4) 单击【模型】按钮，进入【对应分析：模型】对话框。
- ① 在【解的维数】框中，由于本例中样品数和指标(变量)数都较少，故使用系统默认值“2”，即将样品和变量对应地分为两类。
- ② 在【距离度量】栏中选择系统默认的【卡方】距离测度。
- ③ 在【标准化方法】栏中由于在【距离测度】栏中选定了【卡方】距离测度，故在此只能选择系统默认的【行和列均值已删除】(行和列两者被中心化)。
- ④ 在【正态化方法】(应为【常态化方法】)栏中，选择【对称】项。单击【继续】按钮，返回【对应分析】主对话框。
- (5) 单击【统计量】按钮，进入【对应分析：统计量】对话框，选择【对应表】，要求输出对应表；选择【行轮廓表】、【列轮廓表】，输出行、列变量归一化处理表。
- (6) 单击【绘制】按钮，进入【对应分析：图】对话框，只选择【双标图】复选项。
- (7) 单击【确定】按钮，执行运算。
- (8) 输出结果见表 14-18~表 14-21 和图 14-28。

对下文中的表、图是如何生成的进行说明：在 SPSS 20.0 版中，无论是中文版还是英文版，只要对分类变量的值标签用中文进行定义，则在输出窗中得到的对应表、行列归一化处理表均是输出不全的表。

因此，图 14-28 是在 province 分类变量中使用各省中文名称的值标签的情况下得到的，而表 14-18~表 14-21 是在将 province 变量的值标签用拼音的首字母缩写作值标签的情况下得到的。

输出结果包括反映原始数据组成的对应表、行和列的归一化处理表及汇总表。

表 14-18 所示的对应表给出了 10 个省份的 7 种消费支出的观察值、总和，行、列有效边缘值 Active Margin。最右下角的值 9.825 是所有观察值的和。

表 14-18 对应表

对应表						
省份	消费支出					
	食品	衣着	居住	家庭设备及服务消费	医疗保健	交通通讯
SX	.584	.080	.092	.975	.038	.019
NM	.581	.083	.112	.984	.043	.040
LN	.565	.079	.124	.984	.043	.031
JL	.531	.095	.117	.976	.044	.039
HLJ	.555	.073	.143	.984	.052	.026
HN	.655	.097	.095	.983	.022	.019
SC	.640	.072	.117	.990	.034	.017
GZ	.725	.057	.073	.989	.016	.016
GS	.679	.068	.088	.978	.040	.015
QH	.666	.034	.097	.982	.039	.019
有效边际	6.181	.738	1.060	9.825	.372	.241

表 14-19 行归一化处理表

行简要表						
省份	消费支出					
	食品	衣着	居住	家庭设备及服务消费	医疗保健	交通通讯
SX	.599	.082	.095	1.000	.039	.019
NM	.591	.085	.114	1.000	.044	.041
LN	.574	.080	.126	1.000	.044	.032
JL	.544	.098	.120	1.000	.045	.039
HLJ	.564	.074	.146	1.000	.053	.027
HN	.666	.098	.097	1.000	.023	.019
SC	.647	.073	.118	1.000	.034	.018
GZ	.734	.058	.074	1.000	.017	.016
GS	.694	.069	.090	1.000	.041	.016
QH	.678	.034	.099	1.000	.040	.020
质量	.629	.075	.108	.044	.038	.025

表 14-19 所示为对应表中每行观察值除以每行总和的归一化结果。每行的边际都为 1。

表 14-20 所示为对应表中每列观察值除以每列总和的归一化结果。每列的边际都为 1。

表 14-20 列归一化处理表

省份	消费支出					
	食品	衣着	居住	家庭设备及服务消费	医疗保健	交通通讯
SX	.094	.108	.087	.099	.103	.078
NM	.094	.113	.106	.100	.116	.166
LN	.091	.107	.117	.100	.117	.130
JL	.086	.129	.110	.099	.118	.160
HLJ	.090	.099	.135	.100	.140	.109
HN	.106	.131	.090	.100	.060	.077
SC	.104	.098	.110	.101	.090	.072
GZ	.117	.078	.069	.101	.044	.065
GS	.110	.092	.083	.100	.107	.063
QH	.108	.046	.091	.100	.106	.080
有效边际	1.000	1.000	1.000	1.000	1.000	1.000

表 14-21 汇总表

维数	奇异值	惯量	卡方	Sig.	惯量比例		置信奇异值	
					解释	累积	标准差	相关
							2	
1	.133	.018			.657	.657	.309	-.050
2	.070	.005			.180	.837	.294	
3	.050	.002			.091	.928		
4	.036	.001			.047	.976		
5	.024	.001			.021	.996		
6	.010	.000			.004	1.000		
总计		.027	.265	1.000 <sup>a</sup>	1.000	1.000		

a. 54 自由度

表 14-21 所示为汇总表，给出了行与列记分之间的关系，从左到右各列依次为维度、奇异值、惯量、卡方值(即列联表行列独立性卡方检验的卡方值)、显著性水平(即行列独立的零假设下的概率值，值很大说明列联表的行与列之间独立，否则有较强的相关性)、惯量比例。由该表中可见，由于第一维(0.657)、第二维(0.180)的惯量比例和为 83.7%，因此其他维度的重要性可以忽略。

在图 14-28 中，以横轴的“0.0”为中心轴(横轴 0 处的竖线是后加的)，可将变量点和样品点分为两类。

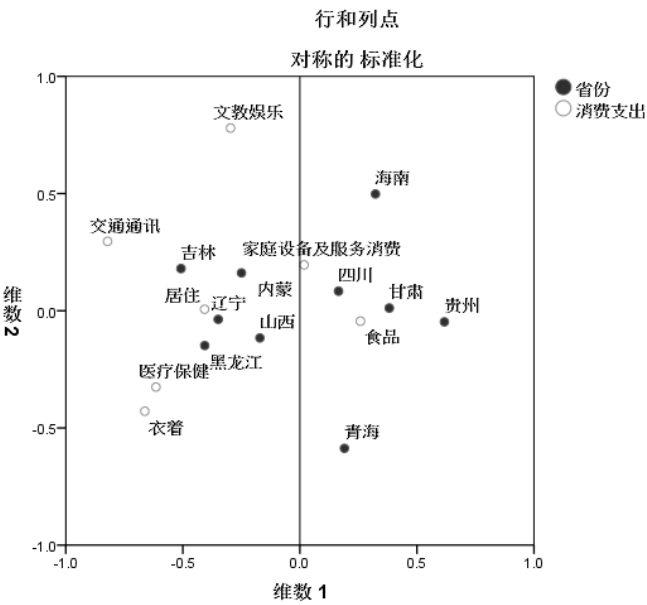


图 14-28 行和列记分图

第一类：变量为衣着、居住、医疗保健，省份有山西、内蒙古、辽宁、吉林、黑龙江，它们位于我国的东部和北部地区，说明这 5 个省份的消费支出结构相似。

第二类：变量为食品、家庭设备及服务消费，省份有四川、贵州、甘肃，它们位于我国的西部和南部地区，说明这 3 个省份的消费支出结构相似。

青海、海南距各种消费类型都较远，较特殊，但这两个省份距离较远，类型又不同。

## 习 题 14

1. 简述主成分分析的基本思想。
2. 用什么统计量衡量主成分中各成分提供的信息量?
3. 一般根据什么确定主成分提取的数量?
4. 简述因子分析的基本思想。
5. 为什么要对初始因子分析结果进行旋转?
6. 简述对应分析的基本思想。对应分析与因子分析有什么不同?
7. 数据文件 **data14-04** 是某医院 3 年中各月的数据, 包括门诊人次、出院人数、病床利用率和周转次数、平均住院天数、治愈或好转率、病死率、诊断符合率、抢救成功率。采用因子分析法探讨综合评价指标。
8. 数据文件 **data14-05** 是 1997 年全国 31 个省、市、自治区按各种经济类型资产占总资产比重(%)的数据, 试对其作对应分析。

# 第 15 章 信度分析与多维尺度分析

在【分析】主菜单中的【度量】分析命令项，如图 15-1 所示，其主要功能是进行信度分析和多维尺度分析，包括的统计功能有 4 项：

- (1) 【可靠性分析】。
  - (2) 【多维展开 (PREFSCAL)】。该方法试图找到一种定量测度方法从而可以直观地研究两个事物之间的关系。
  - (3) 【多维尺度 (PROXSCAL)】。可以作相似性数据和不相相似性数据的分析，比 ALSCAL 功能更强。
  - (4) 【多维尺度 (ALSCAL)】。仅能分析不相相似性数据。
- 本章只介绍可靠性分析和多维尺度 (ALSCAL) 分析。

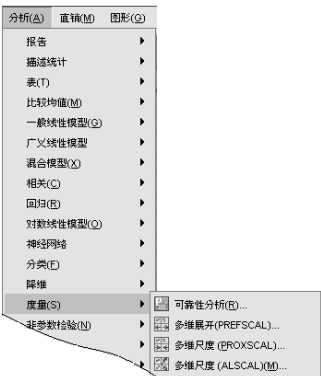


图 15-1 度量分析菜单及其统计命令

## 15.1 信度分析

### 15.1.1 信度分析的概念

#### 1. 什么是信度

信度又叫可靠性，是指测验的可信程度。它主要表现测验结果的一贯性、一致性、再现性和稳定性。一个好的测量工具，对同一事物反复多次测量，其结果应该始终保持不变才可信。比如，用一把尺子测量一批物品，如果今天测量的结果与明天测量的结果不同，那么我们就不会对这把尺子的可信性产生怀疑。信度分析一般在心理学中应用较多，另外在学生考试试卷、社会问卷调查的有效性分析中也会涉及。信度只受随机误差影响，随机误差越大，测验的信度越低。因此，信度也可视为测量结果受随机误差影响的程度。系统误差产生恒定效应，不影响信度。

在测量学中，信度被定义为：一组测量分数的真变异数与总变异数 (实得变异数) 的比率，即

$$r_{xx} = \frac{S_r^2}{S_x^2}$$

式中， $r_{xx}$  为信度系数； $S_r$  为真变异数； $S_x$  为总变异数。

在实际测量中，因为真值是未知的，故信度系数不能由上式直接求出，而只能根据一组实得分数 (测得值) 作出估计。

信度系数是衡量测验好坏的一个重要技术指标，测验的信度系数达到多高才可以接受呢？最理想的情况是  $r = 1$ ，但这是办不到的。大多数学者认为，任何测验或量表的信度系数如果在 0.9 以上，则该测验或量表的信度甚佳；信度系数在 0.8 以上都是可以接受的；如果在 0.7 以上，

则该量表应进行较大修订，但仍不失其价值；如果低于 0.7，量表就需要重新设计了。在心理学中通常可以用已有的同类测验作为比较的标准。一般能力与成就测验的信度系数常在 0.90 以上，性格、兴趣、态度等人格测验的信度系数通常在 0.80~0.85 之间。

2. 相关术语

(1) 量表(scale)。用以测量的准尺。它是一个具有单位和参照点的连续体，将被测量的事物置于该连续体的适当位置，看它离开参照点多少单位的计数，便得到一个测得值。这种连续体就称为量表。量表一般都由一套测验题目构成，其中每一测验题都符合标准化要求，具有一定的分值。

(2) 平行测量。在心理学中能以相同的程度测量同一心理特质的测验。简单地说，就是两个或两个以上的等值测量。同一特质的两个测量，若其测量误差的方差通过检验具有齐次性，就是平行测量，如考试中的 A、B 卷。

(3) 多重记分的测验。相对于二值记分的测验而言，二值记分即答对记分，答错不记分。而对一些由主观性题目(如语文考试中的作文、英语考试中的写作)构成的测验，记分可能是 0 到满分之间的任何一个分数，这种记分方式的测验就称为多重记分的测验。

(4) 项目。或称为题项，即量表或试卷中的题目。

(5) 内在信度。内在信度指的是量表中的一组问题(或整个量表)是否测量的是同一个概念，即这些问题之间的内在一致性如何。如果内在信度系数在 0.8 以上，则可以认为量表有较高的内在一致性。最常用的内在信度系数为克隆巴赫  $\alpha$  系数和折半信度。

(6) 外在信度。指在不同时间进行测量时量表结果的一致性程度。最常用的外在信度指标是重测信度，即用同一问卷在不同时间对同一对象进行重复测量，计算一致程度。

3. 信度估计的方法

由于测验分数的误差来源不同，估计信度的方法也有所不同。关于估计信度的具体方法请参见相关书籍，这里只针对 SPSS 中出现的信度估计方法进行介绍。请注意根据数据类型的不

同选择不同的方法。

表 15-1 累加李克特量表的数据输入结构

编号	问 卷 题 目				
	1	2	3	...	k
1	$x_{11}$	$x_{21}$	$x_{31}$	...	$x_{k1}$
2	$x_{12}$	$x_{22}$	$x_{32}$	...	$x_{k2}$
3	$x_{13}$	$x_{23}$	$x_{33}$	...	$x_{k3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{1n}$	$x_{2n}$	$x_{3n}$	...	$x_{kn}$

(1)  $\alpha$  信度系数。

$\alpha$  信度系数是目前最常用的信度系数。它表明量表中每一题项得分间的一致性。该方法适用于项目多重记分的测验数据或问卷数据，可以用该系数测量累加李克特量表(Likert-type Scale)的信度。累加李克特量表的数据输入格式见表 15-1。

其中， $x_{ij}$  ( $i=1, 2, \dots, k; j=1, 2, \dots, n$ )表示各受试对象第  $i$  个题目的得分。量表共有  $k$  个题目， $n$  名受试对象。 $\alpha$  信度系数公式为

$$\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum_{i=1}^k S_i^2}{S_x^2} \right)$$

式中， $k$  为测验的题目数； $S_i$  为第  $i$  题得分数的方差； $S_x$  为测验总分的方差。

$\alpha$  信度系数可以解释用量表测试某一特质所得分数的变异中，有多大比例是由真分数所决定的，从而反映量表受随机误差影响的程度，即反映出测试的可靠程度。例如， $\alpha=0.90$  时，可以说测试所得分数的 90%的变异是来自真分数的变异，仅有 10%的变异来自随机误差。还可以



把  $\alpha$  信度系数视作相关系数, 它的取值范围为  $0 \sim 1$ , 出现负值是违反可靠性模型的。

用  $\alpha$  信度系数来估计量表的信度时, 应注意  $\alpha$  信度系数与量表题目数量的多少有关。如一个含约 10 个题目的量表, 克隆巴赫  $\alpha$  系数应能达到 0.80 以上。如果题目增加, 克隆巴赫  $\alpha$  系数会随之升高, 题目多于 20 个时, 克隆巴赫  $\alpha$  系数会很容易地升至 0.90 以上; 如果量表的题目减少, 克隆巴赫  $\alpha$  系数会随之降低, 一个 4 个题目的量表, 克隆巴赫  $\alpha$  系数有时可能会低于 0.60 或 0.50。因此, 判断量表信度时, 首先应当了解该量表题目的数量, 然后再以此为基础, 判断克隆巴赫  $\alpha$  系数是否达到了可以接受的水平。

### (2) 分半信度。

任何测验只是所有可能题目中的一份取样, 如果抽取不同的部分, 则可编制很多平行的等值测验, 称为复本(内容、形式相等的测验), 如 A、B 试卷。如果一种测验有两个以上的复本, 根据一群被试接受两个复本测验的得分计算相关系数, 即可得到复本信度, 作可靠性分析。但建立复本是相当困难的, 因此, 在测验没有复本且只能实施一次的情况下, 通常采用分半法估计信度, 即将测验题目分成对等的两半, 根据各人在这两半测验的分数, 计算其相关系数作为信度指标。其计算公式为

$$r_{xx} = \frac{2r_{hh}}{1+r_{hh}}$$

式中,  $r_{hh}$  为两半测验分数的相关系数;  $r_{xx}$  为整个测验的信度估计值。应该注意的是, 如果测验的题目数量较少, 如 10 题以下, 就不适合用这种方法来估计信度。

另外, 分半法的使用基于人为分成两半的测验要等值, 即两半测验的分数具有相同的平均数和标准差。当此条件不能满足时, 就需要采用下面两个公式来估计信度。

#### ① 弗朗那根公式

$$r = 2 \left( 1 - \frac{S_a^2 + S_b^2}{S_x^2} \right)$$

式中,  $S_a^2$  和  $S_b^2$  分别为两半测验分数的方差;  $S_x^2$  为测验总分的方差;  $r$  为信度值。

#### ② 卢伦公式

$$r = 1 - \frac{S_d^2}{S_x^2}$$

式中,  $S_d^2$  为两半测验分数之差的方差;  $S_x^2$  为测验总分数的方差;  $r$  为信度值。

### (3) 库德-理查逊(Cuttman)公式。

倘若一个测验全由二值记分(1, 0 方式记分)的项目所组成,  $\alpha$  信度系数公式中每个项目上的分数方差就会等于该项目上通过率  $p$  与未通过率  $q$  两者的积。库德-理查逊公式为

$$r_{kk} = \frac{k}{k-1} \left( 1 - \frac{\sum p_i q_i}{S_x^2} \right)$$

式中,  $k$  为构成测验的题目数;  $p_i$  为通过第  $i$  题的人数比例;  $q_i$  为未通过第  $i$  题的人数比例;  $S_x^2$  为测验总分的方差。

### (4) 平行测验的信度估计。

对于信度, 也可定义为两平行测验上观察分数间的相关, 即用一个平行测验上某被试的观察分数, 去正确推论另一平行测验上该被试观察分数的能力, 用这种能力值的大小来定义测验的信度。平行测验信度估计的条件是方差具有齐次性, 有时还要求两平行测验的均数相等。

4. 数据要求与假设

(1) 数据要求。用于分析的数据可以是数值型的二分数据，也可以是数值型的有序变量和尺度变量。

(2) 假设。观测值应相互独立，在各项目之间的误差应互不相关。量表是可加的，即各个项目得分相加即为总分数，因此各个项目与总分数是线性相关的。

15.1.2 信度分析过程

(1) 按【分析→度量→可靠性分析】顺序打开如图 15-2 所示的【可靠性分析】主对话框。

(2) 在左侧的源变量框中选择变量进入右侧的【项目】框，作为分析变量。

(3) 在源变量框下面有【模型】下拉列表，用来选择估计信度系数的方法。单击向下箭头，出现 5 种信度估计方法供选择。默认方法是【 $\alpha$ 】信度系数。

①【 $\alpha$ 】。即克隆巴赫  $\alpha$  系数，是内部一致性估计的方法，适用于项目多重记分的测验(主观题)。

②【半分】。也叫分半信度。将测验题分成对等的两半，计算这两半分数的相关系数。

③【Guttman】。适用于测验全由二值(1, 0)方式记分的项目。

④【平行】。是平行测验信度估计的方法，条件是各个项目的方差具有齐次性。

⑤【严格平行】。除了要求各项目方差具有齐次性外，还要求各个项目的均数相等。

(4) 在【刻度标签】框内可以对所计算的信度系数进行标注说明，如输入“自觉性维度的  $\alpha$  系数”。

(5) 单击【统计量】按钮，打开【可靠性分析：统计量】对话框，见图 15-3，在其中选择要输出的统计量。



图 15-2 【可靠性分析】主对话框

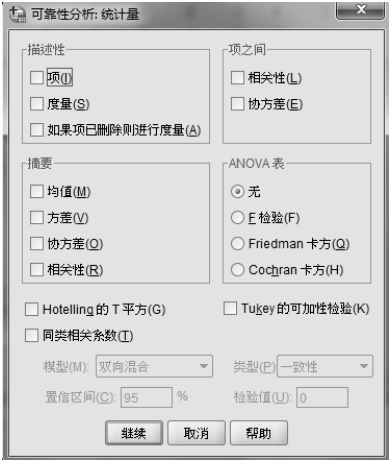


图 15-3 【可靠性分析：统计量】对话框

①【描述性】栏。

- 【项】。计算各项的均数、标准差和样本含量。
- 【度量】。计算量表的均数、标准差和项目数。即将各项目分数汇总得到总分数，将总分数作为变量，求其均数、标准差。
- 【如果项已删除则进行度量】。计算总分减去当前项目得分后的均数、方差等统计量。

②【项之间】栏。

- 【相关性】。计算各项目间的相关系数。
- 【协方差】。计算各项目间的协方差。

③【摘要】栏。计算量表的描述统计量，包括均值、方差、相关系数和协方差。

- 【均值】。对项目均数计算统计量，包括项目均数的平均值、最小值、最大值、极差、最大值与最小值之比和项目均数的方差。
- 【方差】。对项目方差计算统计量，包括项目方差的平均值、最小值、最大值、极差、最大值与最小值之比和项目方差的方差。
- 【协方差】。对项目协方差计算统计量，包括项目协方差的平均值、最小值、最大值、极差、最大值与最小值之比和项目协方差的方差。
- 【相关性】。对项目相关系数计算统计量，包括项目相关系数的平均值、最小值、最大值、极差、最大值与最小值之比和项目相关系数的方差。

④【ANOVA 表】栏。选择方差分析的方法，是对均值相等的检验，即检验各个项目上的得分是否具有的一致性。

- 【无】。不生成方差分析表，即不进行检验。这是系统默认选项。
- 【F 检验】。输出重复测量方差分析表。
- 【Friedman 卡方】。计算 Friedman 卡方值和 Kendall 谐和系数，是对多个配对样本的平均秩之间有无差异的检验。此时 Friedman 卡方检验取代通用的 F 检验。
- 【Cochran 卡方】。显示 Cochran's Q 值。如果项目都是二分变量，选择此项。这时在 ANOVA 表中使用 Q 统计量取代常用的 F 统计量。它也是对多个配对样本的检验。

⑤【Hotelling 的 T 平方】。生成霍特林  $t^2$  统计量，是对所有项目均数相等的零假设的多变量检验。

⑥【Tukey 的可加性检验】。给出量表提高可加性的功效估计值。检验假设是项目间没有交互作用。

⑦【同类相关系数】。输出组内相关系数，同时给出相关系数的置信区间、F 统计量和显著性检验值。选中此项，将激活下面的选项。

- 【模型】下拉列表，选择计算组内相关系数的模型。单击向下箭头，有 3 种选择。
  - A. 【双向混合】。二维混合模型。
  - B. 【双向随机】。二维随机模型。
  - C. 【单向随机】。一维随机模型。
- 【类型】下拉列表。指定组内相关系数 (Intraclass Correlation) 是如何被定义的。
  - A. 【一致性】。表明研究中不关注评分者给出相同分数。
  - B. 【绝对一致】。表明研究者关注评分者给出相同的分数。
- 【置信区间】框。指定置信度，计算置信区间。系统默认值为 95%。
- 【检验值】框。为进行假设检验在此输入一个组内相关系数的假定值。系统默认值是 0，是相关系数为 0 的零假设。

(6) 在主对话框中单击【确定】按钮，提交运行。

### 15.1.3 信度分析实例

【例 1】 本例是心理学中研究运动员意志品质的调查问卷数据，对应数据文件 data15-01。

问卷中有 50 个题目，即 50 个项目。共对 312 人进行了问卷调查。根据数据资料进行项目分析(即对问卷作因子分析，有关因子分析的内容参见第 14 章)后，删除第 7、8、14、28、29、35、36、37、38、40、43、48 题，并将剩余的 38 题根据项目分析的结果分为 5 个维度，所包括的项目如下。

自觉性维度：x1、x2、x4、x10、x13、x39、x41、x45，共 8 题。

果断性维度：x25、x31、x32、x34、x42、x44、x47、x49、x50，共 9 题。

自制力维度：x3、x6、x15、x17、x18、x21，共 6 题。

坚韧性维度：x5、x9、x11、x12、x16、x20、x23、x24、x26、x30、x46，共 11 题。

主动性维度：x19、x22、x27、x33，共 4 题。

表 15-2 所示为运动员意志品质评价量表(预测版)的格式。

现在要检验问卷的内部一致性，即进行信度分析。用 SPSS 中的 Scale 信度分析功能求得  $\alpha$  系数，就能说明该问卷中的 38 个变量的内部一致性结构。具体操作步骤如下。

表 15-2 运动员意志品质评价量表(预测版)

测 试 内 容	评 定 等 级				
	完全不符合	不太符合	说不清楚	比较符合	完全符合
1. 只要是一件有意义的事，我就会去做	1	2	3	4	5
2. 在一天的训练中，即使是重复同一动作，我也会一丝不苟地完成	1	2	3	4	5
3. 一件事完成以后，我经常后悔自己为何不早下决心	1	2	3	4	5
...	...	...	...	...	...
15. 对困难的任务我会想尽办法完成	1	2	3	4	5
16. 在以往比赛中，因为犹豫我错过了许多机会	1	2	3	4	5
...	...	...	...	...	...
24. 我决定做一件事时，常常是说干就干，决不拖拉或让它落空	1	2	3	4	5
25. 我常常为做决定犯难	1	2	3	4	5
...	...	...	...	...	...
27.遇到棘手的事情我常常举棋不定，拿不出主意	1	2	3	4	5
...	...	...	...	...	...
32.我主动找过教练商量下一步的练习计划	1	2	3	4	5
33.如果见到有人落水，我会马上去救他	1	2	3	4	5

- (1) 建立数据集，见数据文件 data15-01。x1~x50 是问卷的题目，即项目。
- (2) 按【分析→度量→可靠性分析】顺序打开【可靠性分析】主对话框。
- (3) 在主对话框的源变量框中，选中自觉性维度题项变量 x1、x2、x4、x10、x13、x39、x41、x45 进入【项目】框。
- (4) 在源变量框下面的【模型】下拉列表中选择信度估计方法。本题选用系统默认的【 $\alpha$ 】系数。
- (5) 在【刻度标签】框内输入“自觉性维度的信度系数”。
- (6) 单击【确定】按钮，提交运行，计算自觉性维度的信度系数  $\alpha$ 。
- 重复第(2)步，在主对话框中单击【重置】按钮，通过第(3)、(4)、(5)、(6)步分别对果断性、自制力、坚韧性、主动性维度计算信度系数。
- (7) 主动性维度的信度系数求出后，在主对话框中将 38 个项目全部送入【项目】框，信度估计法还用系统默认的【 $\alpha$ 】系数。单击【确定】按钮提交运行，输出结果整理于表 15-3 中。
- 需要注意的是，如果问卷中有反向题(如果正向题给予 1、2、3、4、5 分，而反向题给予

的是 5、4、3、2、1 分)，则需要将其转换。

可以在输入数据时就直接将反向题进行转换。

如果输入数据时未进行转换，可以单击【转换→重新编码为相同变量】完成，详见第 2.3.3 节。本例的数据已经对反向题进行了转换。

(8) 输出结果解释。

表 15-3 中，5 个维度的信度系数分别为 0.144、0.441、0.271、0.519、0.042，而总量表的信度系数是 0.636。

表 15-3 量表及各维度的信度系数表

维度 <sup>①</sup>	信度分析 <sup>②</sup>		
	样本量(N) <sup>③</sup>	项数 <sup>④</sup>	α 系数 <sup>⑤</sup>
自觉性维度 <sup>⑥</sup>	312 <sup>⑦</sup>	8 <sup>⑧</sup>	0.144 <sup>⑨</sup>
果断性维度 <sup>⑥</sup>	312 <sup>⑦</sup>	9 <sup>⑧</sup>	0.441 <sup>⑨</sup>
自制力维度 <sup>⑥</sup>	312 <sup>⑦</sup>	6 <sup>⑧</sup>	0.271 <sup>⑨</sup>
坚韧性维度 <sup>⑥</sup>	312 <sup>⑦</sup>	11 <sup>⑧</sup>	0.519 <sup>⑨</sup>
主动性维度 <sup>⑥</sup>	312 <sup>⑦</sup>	4 <sup>⑧</sup>	0.042 <sup>⑨</sup>
量表 <sup>⑥</sup>	312 <sup>⑦</sup>	38 <sup>⑧</sup>	0.636 <sup>⑨</sup>

5 个维度的信度系数都偏低，需要进行问卷的修改。此外，总量表的信度系数是 0.636，代表该量表的信度一般。如果要提高信度系数，可以从 5 个维度中项目内容词句进行修饰、修改，如果时间允许，可增删项目，再让这 312 名受试者测试一次；如果时间不允许，应在研究论文中加以说明，可作为今后研究的方向。

## 15.2 多维尺度分析(ALSCAL)

### 15.2.1 多维尺度分析的功能与数据要求

多维尺度分析(Multidimensional Scaling)是市场调查、分析数据的统计方法之一。通过多维尺度分析，可以将消费者对商品相似性的判断产生一张能够看出这些商品间相关性的图形。例如，有十个百货商场，让消费者排列出对这些百货商场两两间相似的感知程度，根据这些数据，用多维尺度分析可以判断消费者认为哪些商场是相似的，从而可以判断竞争对手。

(1) 数据要求。如果数据为不相似性数据，它们必须为数值型数据或者是使用相同计量单位计量的数据。如果数据为多元变量数据，则数据可以是等间隔数据、二分数据或者是计数数据。注意应该保持数据量度单位的一致性，否则将会影响到分析结果。如果不能避免这种情况的出现，必须对数据进行标准化(在此分析过程中，可以自动解决)。

(2) 假设。多维尺度分析没有严格的假设要求，但在选择测量水平时应该十分小心。

### 15.2.2 多维尺度分析过程

- (1) 按【分析→度量→多维尺度(ALSCAL)】顺序打开【多维尺度】主对话框，见图 15-4。
- (2) 在左侧的变量表中选择变量，单击向右箭头，将变量送入【变量】框中。

【变量】框下有【单个矩阵】框，在此可输入一个变量进行分组。程序会为每一组分别计算距离矩阵，同时无论是否在【模型】对话框中选择了个体差异距离模型，都会计算一个复本或加权距离模型。

- (3) 【距离】栏。
  - ①【数据为距离数据】。当数据视图窗口中的数据是一个或多个不相似性矩阵时选择此项。矩阵中的元素显示行和列两两配对的不相似程度。在【形状】按钮旁边显示的是当前选项。单击【形状】按钮，打开【多维尺度：数据形状】对话框，见图 15-5。
  - 【正对称】。方形对称结构。行、列代表相同的项目，且在上、下三角中相应的值相等。

例如 A、B 两个项的相似性，A 与 B 和 B 与 A 的相似性是一样的，矩阵中上下三角对应位置上的值是相等的。

- 【正不对称】。方形但不对称结构。行、列代表相同的项目，但上三角和下三角中相应的值是不相等的。如两个事物 A 与 B 比较和 B 与 A 比较所得分数不同。
- 【矩形】矩形结构。在【行数】框中输入行数。在矩阵中，行、列数据代表不同项目集。SPSS 把有序排列的数据文件当作矩形矩阵。如果数据中包含两个以上的矩形矩阵，一定要设定每个矩阵的行数。此数值必须大于等于 4，并且能够将矩阵中的行数整除(即各矩阵的行数应当相同)。



图 15-4 【多维尺度】主对话框

② 【从数据创建距离】。根据数据生成距离阵。在【度量】按钮旁边显示的是当前选项。单击【度量】按钮，打开【多维尺度：从数据创建度量】对话框，见图 15-6。

- 【度量标准】栏。可在此处选择用于分析的不相似性度量方法，方法说明见附录 A。
- 【创建距离矩阵】栏。
- A. 【变量间】。计算一对对变量之间的不相似性距离矩阵。
- B. 【个案间】。计算两两观测量之间的不相似性距离矩阵。
- 【转换值】栏。进行标准化转换，方法说明或算法见附录 A。

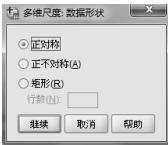


图 15-5 【多维尺度：数据形状】对话框

(4) 在主对话框中单击【模型】按钮，进入【多维尺度：模型】对话框，见图 15-7。在该对话框中确定数据和模型的类型。多维尺度分析的正确估计依赖于数据和模型。



图 15-6 【多维尺度：从数据中创建度量】对话框

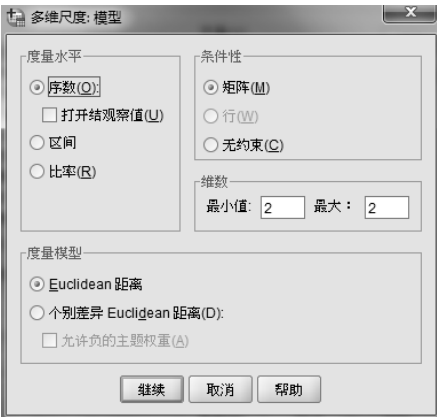


图 15-7 【多维尺度：模型】对话框

- ① 【度量水平】栏。在该栏指定测度水平。有 3 个单选项：
  - 【序数】。有序测度的数据。若是有序分类数据，使用 Kruskal 最小平方单调转换。选择【打开结观察值】，将对有相同分数的观测值赋予不同的秩。
  - 【区间】。数据是以间隔测度或定量的数据。

● **【比率】**。数据是比例测度或定量的数据。

② **【条件性】** 栏。指定模型类型，有 3 个单选项：

● **【矩阵】**。如果只有一个矩阵或每个矩阵代表不同的受试者时，选择此项。

● **【行】**。只有行数据进行比较时有意义，选择此项，且只适用于不对称或矩形矩阵。

● **【无约束】**。矩阵内所有数值的比较都有意义时选择此项。

③ **【维数】** 栏。用来指定多维尺度分析的维度。默认产生二维解。在 **【最小值】** 框中输入最少维度数，在 **【最大值】** 框中输入最多维度数。一般可选择计算 1~6 维度的解。为获得唯一解，在最小和最大维数栏中输入相同的数值。对于加权模型，**【最小值】** 栏中至少应是 2。

④ **【度量模型】** 栏。指定尺度模型，有两个单选项：

● **【Euclidean 距离】**。欧几里得模型可以应用于任何类型的矩阵分析中。如果数据中只包含一个矩阵，那么将进行 CMDS 典型多维尺度分析；如果包含两个以上的矩阵，则进行 RMDS 重复多维尺度分析。

● **【个别差异 Euclidean 距离】**。加权个体差异欧几里得距离模型 (WMDS)。该模型需要两个或以上的矩阵。

(5) 单击 **【选项】** 按钮，进入 **【多维尺度：选项】** 对话框，

见图 15-8。

① **【输出】** 栏。在该栏中选择输出项。

● **【组图】**。多维尺度分析图。这个图在多维尺度分析中非常重要。可以利用这个图对每一维寻找散点间相关性的合理的解释。

● **【个别主题图】**。对有序分类数据或模型中指定矩阵的数据显示每一个受试者的图形，而对模型中指定行的数据无效。

● **【数据矩阵】**。显示每一个受试者的数据矩阵。

● **【模型和选项摘要】**。输出所有选项的基本信息，包括数据选项、模型选项、输出选项和迭代判据选项等信息。

② **【标准】** 栏。设置迭代停止的判据，有 3 个选项：

● **【S 应力收敛性】** 框。单调收敛准则，系统默认在拟合距离模型过程中计算拟合劣度指标 S-stress。当从一个迭代到下一个迭代的 S-stress 变化量 (即拟合的改善量) 等于或小于 0.001 时，迭代停止。为了提高解的精度，可以输入一个比以前设置值小的正值。如果输入 “0”，只进行 30 步的迭代。

● **【最小 s 应力值】** 框。系统默认收敛值为 0.005 时迭代停止。如果要继续进行迭代，输入一个比默认值更小的数值。如果输入的数值比默认值大，迭代次数会减少。该值要大于 0，小于或等于 1。

● **【最大迭代】** 框。用最大迭代次数作为迭代停止的判据。当最大迭代次数等于设置值时迭代停止。系统默认值是 30。如果输入值比默认值大，可增加分析的精度，但计算时间也会增加。

③ **【将小于…的距离看作缺失值】** 框。系统默认将距离小于 0 的值作为缺失值。用户可以指定  $n$  值，则系统把小于  $n$  的值作为缺失值处理。

(5) 在主对话框中单击 **【确定】** 按钮，执行操作，系统将输出多维尺度分析结果。

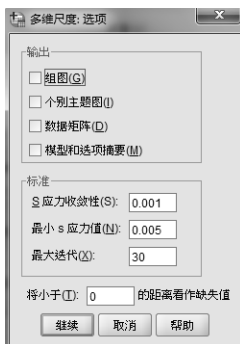


图 15-8 **【多维尺度：选项】** 对话框

15.2.3 多维尺度分析实例

**【例 2】** 本例使用《SAS 系统与市场调查数据分析》(高慧璇等编著)一书中的例题数据。该数据是假设 7 名受试者按照 1~7 的尺度(1 表示非常相似, 7 表示非常不相似)排列出一些饮料间两两相似的感知程度。这些饮料作为变量包括 milk(牛奶)、coffee(咖啡)、tea(茶)、soda(苏打水)、juice(果汁)、botwater(矿泉水)、beer(啤酒)、wine(葡萄酒)。要求受试者给出这些饮料的两两相似的感知程度, 共有 28 种可能 $[n(n-1)/2]$ 。用此数据分析哪些饮料消费者认为是相似的。该分析可以使用多维尺度分析(ALSCAL)方法完成。具体步骤如下:

(1) 打开数据文件 data15-02。因为本例的数据矩阵是对称的, 例如牛奶与咖啡间的距离和咖啡与牛奶间的距离一样, 所以可以作成三角矩阵。sub 变量为受试者编号。每个受试者对 7 种饮料两两比较, 根据它们之间的相似度打分, 7 分制, 所给分值越大表明相似程度越高, 则定义为相似数据; 所给分值越小表明相似程度越高, 则定义为不相似数据。例如, 图 15-9 所示的数据就是不相似数据, 第一个受试者认为牛奶与牛奶非常相似, 两者的相似度打分为 1; 咖啡与牛奶不相似, 认为两者的相似度为 6, 依此类推。每个受试者的数据是一个矩阵, 其数据结构见图 15-9。

sub	sort	milk	coffee	tea	soda	juice	botwater	beer	wine
sub1	milk	1	6	6	7	7	7	7	7
sub1	coffee	6	1	1	7	7	7	7	6
sub1	tea	6	1	1	7	5	4	7	5
sub1	soda	7	7	7	1	5	3	5	4
sub1	juice	7	7	5	5	1	5	3	2
sub1	botwater	7	7	4	3	5	1	6	6
sub1	beer	7	7	7	5	3	6	1	1
sub1	wine	7	6	5	4	2	6	1	1
sub2	milk	1	5	7	7	7	7	3	7
sub2	coffee	5	1	6	7	7	6	7	7
sub2	tea	7	6	1	6	4	4	3	7
sub2	soda	7	7	6	1	7	7	7	4
sub2	juice	7	7	4	7	1	4	6	4
sub2	botwater	7	6	4	7	4	1	7	7
sub2	beer	3	7	3	7	6	7	1	5
sub2	wine	7	7	7	4	4	7	5	1

图 15-9 data15-03 数据文件结构

- (2) 按【分析→度量→多维尺度(ALSCAL)】顺序打开【多维尺度】主对话框。
- (3) 选中分析变量 milk、coffee、tea、soda、juice、botwater、beer、wine, 送入右侧【变量】框。注意输入分析变量的顺序一定要与数据文件中的顺序一致。
- (4) 在【距离】栏中选中【数据为距离数据】项, 单击【形状】按钮, 打开【多维尺度: 数据形状】对话框。
- (5) 在【多维尺度: 数据形状】对话框中选中【正对称】项, 因为本例数据的行与列项目相同, 上三角与下三角的值是相同的。
- (6) 在主对话框中单击【模型】按钮, 打开【多维尺度: 模型】对话框。
- ①【度量水平】栏。因为用 1~7 给饮料的相似度评分, 所以选择【序数】。
- ②【度量模型】栏。选【Euclidean 距离】项, 要求拟合欧几里得距离模型。
- ③【条件性】栏。选【矩阵】项。因为每个矩阵代表一个被试的答案。计算二维解, 故在【维数】栏的【最小值】框和【最大值】框内均输入“2”。



- (7) 在主对话框中单击【选项】按钮，打开【多维尺度：选项】对话框。
- ①【输出】栏。选【组图】项，作多维尺度分析图。
- ②【标准】栏，使用系统默认值。
- (8) 输出结果见图 15-10~图 15-13 和表 15-4。最关注的是多维尺度分析图(有的参考书称之为共用感知图)。
- (9) 结果解释。

图 15-10 给出了二维解决方案迭代过程。在 Criteria 栏指定的迭代最大数为 30，但当拟合劣度 S-Stress 的改善值小于 0.001 时迭代终止。本例迭代到第四步时 S-Stress 的改善值是 0.00062，其值小于 0.001，迭代过程结束。

图 15-11 给出了 Stress 和 RSQ 值。RSQ 即  $R^2$ ，它是拟合优度指标，数值越接近 1，表明模型拟合越好；Stress 是拟合劣度指标，百分比值越大说明模型拟合越差。表 15-4 给出了 Stress 大小与拟合好坏的一个参考。

本例的 Stress 值为 0.30437(30.4%)，RSQ 值为 0.37281，表明模型拟合的不好。解决方法，一个是用近似多维尺度分析 PROXSCAL 方法，另一个是再增加受试者。

图 15-12 给出了二维导出构形表，表中的数值是用在多维尺度分析图的坐标值。

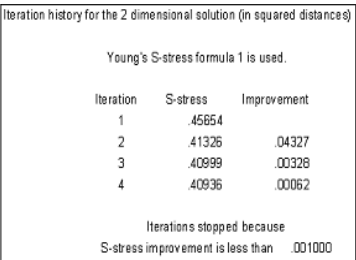


图 15-10 二维解决方案迭代过程

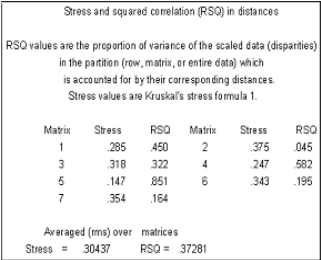


图 15-11 Stress 和 RSQ 值

表 15-4 拟合量度值评价

Stress (%)	拟合度
20	差
10	一般
5	好
2.5	较好

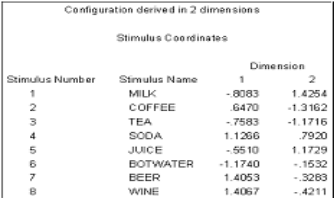


图 15-12 二维导出构形表

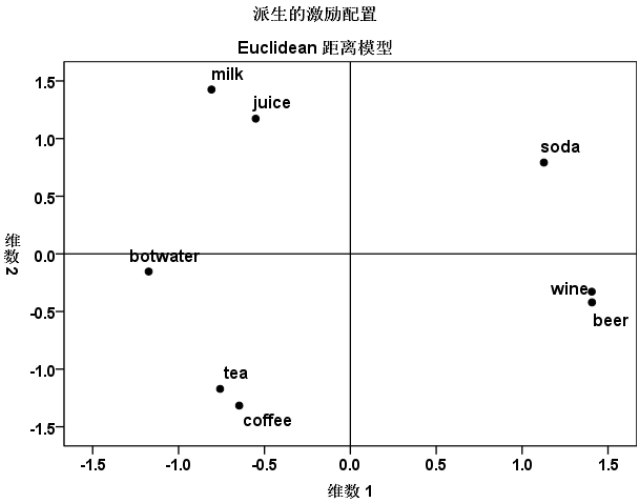


图 15-13 多维尺度分析图

图 15-13 所示为多维尺度分析图，是进行多维尺度分析最关注的结果。从图中可解释的内容包括对图形的每一维寻找散点间相关性的合理解释。图中包括三组聚焦点，这意味着消

费者认为彼此相似的这些产品，即咖啡和茶是相似的，果汁和牛奶是相似的，啤酒和葡萄酒是相似的，说明这些相似饮料在市场占有率上彼此有竞争。另外，从垂直维数 2 看，可将 7 种饮料分为两类，牛奶、果汁、苏打水和矿泉水属于营养型饮料，啤酒、葡萄酒、咖啡和茶属于提神型饮料。

## 习 题 15

1. 信度分析中用哪些指标可以反映问卷可靠性？如果要了解问卷中的某一个维度的可靠性程度如何，应怎么做？
2. 反映量表内部一致性高低的克隆巴赫( $\alpha$ )系数与量表题目的数量有关吗？
3. 什么是相似数据？什么是不相似数据？多维尺度分析的目的是什么？
4. 数据文件 data15-03 中是一个受试者对牙膏认识的数据。试进行不相似数据的多维尺度分析。

# 第16章 结合分析

## 16.1 结合分析概述

### 1. 结合分析的概念

结合分析是测度消费者对产品属性各侧面或售后服务等的偏好的一种技术。

在市场调查中经常想了解顾客对产品的偏好，以作为产品销售策略制定的依据或者作为新产品研制决策的依据。

每个作为商品的产品都是由一系列的属性和服务构成的。例如，一台计算机的属性有 CPU、显示器、内存、硬盘、品牌、价格以及售后服务等，每个属性描述了一台计算机的一个侧面。在计算机技术发展的一个特定时期内，每个属性都有几个技术指标。例如，CPU 有单核、双核、三核、四核之分，显示器有 14in、15in、17in、21in 等，硬盘有 100GB、500GB、800GB 及 1TB 等，内存有 256MB、512MB、1GB、2GB 等。每个属性就是顾客购买决策所考虑的因素，每个属性的每个技术指标就是这个因素的一个水平。

市场研究中的顾客偏好调查分析要求被访者对各种属性水平的组合给出自己的偏好得分，或者将各种组合排序，给出各种组合的秩。这些得分或者排序后的秩分就是对顾客偏好的测度。依据这些数据可分析判断得出顾客偏好的结论。

### 2. 结合分析的步骤

这里阐述的步骤与 SPSS 的结合分析程序有关。

(1) 分析产品的属性。确定每个属性的水平数和水平的具体内容。在统计分析中也称属性为因素。

应选择课题研究的主要因素。选择有代表性的重要的属性是偏好分析的重要环节。属性的数量应该尽量精简。每个属性的水平数也应在达到课题要求的前提下尽量少。

#### (2) 试验设计。

将选择的属性(因素)水平组合成试验组，每个组的因素水平组合称为产品的一个侧面。它是要呈现在被访者面前，供被访者评价的。为减小误差，节省人力、物力、时间，使调查更加有效，通常采用正交设计。可以使用 SPSS 的正交设计程序产生要求数目的侧面，也可以由读者输入形成设计文件。

SPSS 的正交设计程序产生设计文件供调查使用，同时也是结合分析的必要数据。

#### (3) 根据设计打印调查卡片。

#### (4) 运用各种调查方法取得数据。调查取得的数据有两种：

- ① 要求被访者对所设计的侧面排秩，如最喜欢的秩为 1，次之的秩为 2，依此类推。
- ② 要求被访者为所设计的侧面打分，如最喜欢的分数为 100，最不喜欢的分数最低。

(5) 程序设计与运行。SPSS 中没有窗口式的结合分析程序, 必须使用 SPSS 语句进行程序设计, 运行设计的程序分析调查得到的数据, 在输出窗得到输出结果。

(6) 根据输出结果选择顾客最偏爱的产品属性组合, 作为开发新产品的决策依据, 或制定销售策略的依据。

### 3. 本章用到的术语

(1) 侧面。是指所研究的产品属性(因素)水平的组合, 在正交设计中产生。

(2) 全概念侧面。是指能代表各种属性的全部组合。正交设计结果中的侧面可称为全概念侧面, 用此进行偏好调查, 分析结果将是可信的决策依据。

(3) 试验侧面。即要打印成卡片或出现在调查问卷中、由被访者评价的侧面。它是由正交设计形成的。

(4) 保留侧面。是正交设计侧面以外的, 为进行对估计效应有效性的检验而建立的侧面。保留侧面由另一个随机设计产生, 不是由正交设计产生的。

(5) 模拟侧面。由读者输入的侧面。

(6) 设计文件。由正交设计过程生成或者由用户输入, 符合正交性的数据文件。文件中的变量就是课题确定的感兴趣的因素, 是所研究产品的一个属性。而观测量是由各变量水平值组成的产品的侧面。它是一个各因素水平的组合。

设计文件还可以包括保留侧面和模拟侧面。这两个侧面应该由一个特殊变量 Status\_标识, 保留侧面和模拟侧面的 Status\_值分别为 1、2; 试验侧面的 STATUS\_=0。

## 16.2 正交试验设计

### 16.2.1 试验设计中的问题

众所周知, 在调查中, 产品的属性数和各属性的水平数不能太多, 否则其组合数就会很大, 以至于调查和获取数据简直就是不可能完成的任务。例如, 有 2 个属性, 每个属性取 3 个水平, 就有 9 种组合; 如果每个属性有 5 个水平, 就会有 25 种组合; 如果有 5 个因素, 每个因素有 3 个水平, 则组合数是  $243 (3 \times 3 \times 3 \times 3 \times 3)$ , 要求顾客对 243 种产品打分、排序肯定得不到很好的结果。因此要选择有代表性的属性和水平, 而且要有效地减少调查中呈现在调查对象面前的组合。

因此进行试验设计时要考虑:

(1) 当要调查的产品属性(因素)不止一个, 而且每个属性的水平也不止一个时, 要合理安排各个属性水平组合, 以便降低由于被访者对组合理解的差异所引起的误差。

(2) 以最少的属性(因素)水平组合数进行调查, 得到可靠的结论。

(3) 节省人力、物力、财力和时间。

(4) 便于使用软件进行结合分析, 提高调查对顾客偏爱估计的准确性。

### 16.2.2 正交试验设计的思路

为简化问题, 下面以 3 因素 2 水平的试验设计为例进行介绍。

为调查酸奶饮品的顾客偏爱, 选择酸奶的品牌、直接原料、附加成分为产品调查的侧面, 每个侧面选择 2 个水平。要调查哪种水平组合是顾客最爱的。

3 个因素用大写字母表示为 A(品牌)、B(直接原料)、C(成分), 可以看成是一个三维坐标系, 见图 16-1。

各因素的水平序列号用跟在因素字母后的阿拉伯数字表示:

- (1) A 因素。品牌的 2 个水平为 A1(三元)、A2(伊利);
- (2) B 因素。直接原料的 2 个水平为 B1(鲜牛奶)、B2(纯牛奶);
- (3) C 因素。附加成分的 2 个水平为 C1(VAD)、C2(高钙)。

全面试验, 即各因素的各水平全部组合一次, 有  $2^3=8$  次试验, 如表 16-1 所示。这些组合可以用正方体的 8 个顶点表示, 分别是  $A_1B_1C_1$ ,  $A_2B_1C_1$ ,  $A_1B_1C_2$ ,  $A_2B_1C_2$ ,  $A_1B_2C_1$ ,  $A_2B_2C_1$ ,  $A_1B_2C_2$ ,  $A_2B_2C_2$ , 见图 16-2。

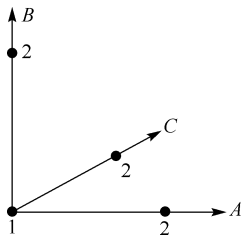


图 16-1 3 个因素各有 2 个水平

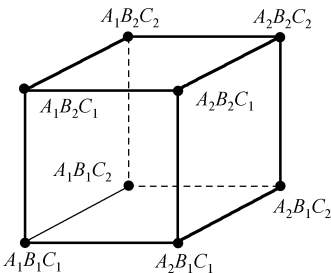


图 16-2 全面试验组合示意

表 16-1 全面试验组合表

		A <sub>1</sub>	A <sub>2</sub>
B <sub>1</sub>	C <sub>1</sub>	A <sub>1</sub> B <sub>1</sub> C <sub>1</sub>	A <sub>2</sub> B <sub>1</sub> C <sub>1</sub>
	C <sub>2</sub>	A <sub>1</sub> B <sub>1</sub> C <sub>2</sub>	A <sub>2</sub> B <sub>1</sub> C <sub>2</sub>
B <sub>2</sub>	C <sub>1</sub>	A <sub>1</sub> B <sub>2</sub> C <sub>1</sub>	A <sub>2</sub> B <sub>2</sub> C <sub>1</sub>
	C <sub>2</sub>	A <sub>1</sub> B <sub>2</sub> C <sub>2</sub>	A <sub>2</sub> B <sub>2</sub> C <sub>2</sub>

再看图 16-3, 只取 4 种水平组合就可以代替上述 8 种全面试验组合, 简化成 4 个试验:  $A_1B_1C_1$ 、 $A_2B_1C_2$ 、 $A_1B_2C_2$ 、 $A_2B_2C_1$ , 见表 16-2。

为什么 4 次试验可以代替 8 次试验呢?

(1) 观察正方体, 3 个因素的每个水平都均匀地包含在这 4 个组合中。选中的 4 个组合均匀地分布在正方体中, 每个面都有 2 个点, 每个线都有一个点, 分布均匀。

(2) 观察正交表 16-3, 具有下列特点:

- ① 每个因素(列)的每个水平都出现, 且出现的次数相同: 1 水平出现两次, 2 水平出现两次。

表 16-2 4 个顶点的组合

	A <sub>1</sub>	A <sub>2</sub>
B <sub>1</sub>	C <sub>1</sub>	C <sub>2</sub>
B <sub>2</sub>	C <sub>2</sub>	C <sub>1</sub>

表 16-3 正交设计结果表

列号 试验号	A	B	C
1	1	1	1
2	2	1	2
3	1	2	2
4	2	2	1

② 任意 2 个因素(任意两列)的水平数据对是相同的。表 16-3 中, A、B 两列的水平搭配是(1,1)、(2,1)、(1,2)、(2,2); A、C 两列的水平数据对是(1,1)、(2,2)、(1,2)、(2,1)与 A、B 两列是相同的, 只是顺序不同。所以因素水平的搭配是均匀的。

具有上述特点的试验设计表就称为正交表。

以上两个性质称为正交表的正交性。这个性质使得正交表在使用部分组合进行试验时具有以下特点:

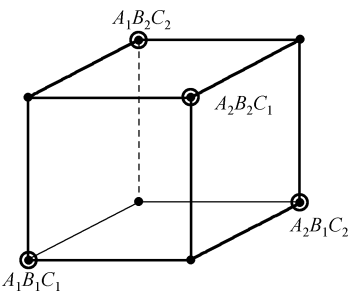


图 16-3 正交设计示意图

(1) 正交表中列出的试验组合能很好地代表全面的试验组合。  
第一个性质各因素各水平在每列中都出现相同的次数，保证了这些试验组合对全面试验的代表性。

第二个性质使任意两个因素间的组合为全面的试验组合，从而保证了使部分试验找到的最优组合与全面试验的结果趋势一致。

(2) 试验组合均衡地分布在全面试验组合之中，见图 16-3。

(3) 正交性使得任一因素各水平的试验条件相同。这就保证了在每列因素各水平的效果中，最大限度地排除了其他因素的

干扰，从而可以综合比较该因素不同水平对试验指标的影响情况。

根据上述正交设计的结果，在调查问卷中呈现在被访者面前的是这样的牛奶的属性组合：

- ① 三元牌 VAD 鲜牛奶；
- ② 伊利牌 高钙 鲜牛奶；
- ③ 三元牌 高钙 纯牛奶；
- ④ 伊利牌 VAD 纯牛奶。

16.2.3 正交试验设计过程

打开一个空数据窗口，设计结构可以占据这个数据窗，也可以保存到存储设备中。

按【数据→正交设计→生成】顺序单击菜单项，打开【生成正交设计】主对话框，见图 16-4 和图 16-5。

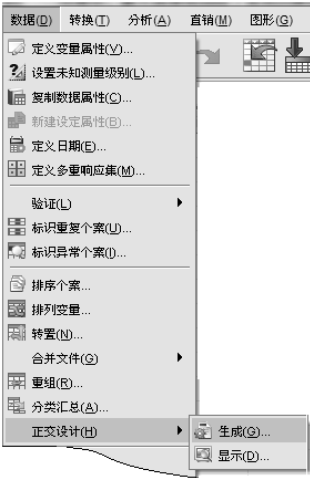


图 16-4 正交设计菜单选择



图 16-5 【生成正交设计】主对话框

(1) 定义因素和因素水平。步骤如下：

- ① 在【因子名称】栏中输入因素变量名，必须是合法的 SPSS 变量名，但不能用 Status\_ 和 Card\_ 作为因素变量名。
- ② 在【因子标签】栏中输入因素变量的标签。
- ③ 单击【添加】按钮将因素变量名及其标签添加到大矩形框中。显示格式为  
因素变量名 ‘标签’ (?)

可以重复上述三步操作，定义若干个因素变量。

如果要修改因素变量名和标签的定义，先选择它，因素变量名和标签重新返回【因子名称】栏和【因子标签】栏。在这两个栏中修改后，单击被激活的【更改】按钮，修改后的因素变量名和变量标签即显示在大矩形框中。如果要删除已经定义的因素变量及其标签，只要在选中后单击【删除】按钮即可。

④ 从大矩形框中选择一个因素变量，激活【定义值】按钮，单击打开相应的二级对话框，见图 16-6。

⑤ 在二级对话框中定义因素变量的值和值标签。

- 如果水平较多，且水平值是从 1 开始的，则可以利用【自动填充】栏中的自动功能，自动填入因素水平值。方法是将水平数输入到【从 1 至】框中，单击【填充】按钮。例如，图 16-6 中输入“3”，单击【填充】按钮后，在【值】栏中自动填入 1、2、3 三个水平值。
- 将各水平值的含义作为标签输入对应水平值后边的【标签】框中。
- 单击【继续】按钮返回主对话框。重复步骤④、⑤，将所有因素变量的值标签定义工作完成。

(2) 定义设计结果保存方式。在主对话框的【数据文件】栏内根据保存要求，选择保存方式。

① 【创建新数据集】。把设计结果保存到一个数据文件中。在【数据集名称】框中输入文件名即可。生成的数据集占据当前数据窗口。由于没有保存到存储设备中，在窗口标题栏中的数据集名称前出现“未标题 n”字样

② 如果想保存该设计文件，选择【创建新数据文件】，单击【文件】按钮，打开【生成 Orthogonal 设计：生成文件说明】对话框。指定保存位置、文件类型和文件名，单击【保存】按钮，返回主对话框，否则将以默认的文件名 ORTHO.sav 保存到默认的位置。



图 16-6 【生成设计：定义值】对话框

(3) 选中【将随机数初始值重置为】复选项后，在其后的框中重新为随机数种子指定一个值。在生成正交设计过程中，要通过随机数种子，产生随机数。相同的随机数种子产生相同的设计结果，因此不同的设计要设置不同的种子值。必须在生成第一个设计之前设置该种子值，种子值可以是 1~2 000 000 000 中的任意整数。在一个 SPSS 执行周期中，如果想生成几个相同的随机数集，并在后续的设计生成时将种子值再次设置成相同的值。

(4) 单击【选项】按钮，打开相应的对话框，见图 16-7。

① 【生成的最小个案数】框中。指定一个计划设计要生成的最少观测数，即指定一个试验设计的最少试验次数。选择一个正整数，要小于等于根据所有因素水平可能的组合构成的观测总数，也就是要小于等于全模型的试验次数。

如果不明确指定要生成的观测次最小数，则自动生成对正交设计必要的数量的观测。如果正交设计过程不能产生至少是大致所要求的最少观测数，就将产生符合所指定的因素和水平数的最大数。注意：该设计没有必要包括确切的指定的观测数，但使用这个值作为最小值在正交设计中生成更合适的、可能是最小观测数。例如，A 因素有 3 个水平、B 因素有 2 个水平，C 因素有 2 个水平，总组合数是 12。设置要生成的最小观测数必须小于或等于 12。

② 【延续个案】栏。设置有关产生除正规设计的观测(个案)以外的保留观测(个案)的数量。

- **【延续个案数】**。设置除正规设计外的观测(个案)数。但结合分析过程估计效应时不使用这些额外的观测。在被激活的框中输入一个正整数,该数值必须小于等于由因素水平组合决定的观测量总数。
- **【与其他个案随机混合】**。输出结果随机地将保留观测与试验观测混合。如果不选择此项,保留观测在数据文件中出现在试验观测后面。

如果选择了**【延续个案数】**,则可在其后的矩形框中给出这个保留个案数,也可以选择**【与其他个案随机混合】**。



图 16-7 **【生成正交设计: 选项】**对话框

- 单击**【继续】**按钮,返回主对话框。

(5) 在主对话框中单击**【确定】**按钮,提交系统执行,输出的设计结果保存到指定位置,并在输出窗给出可能条件组合的设计结果。

16.2.4 正交试验设计实例

**【例 1】** 要求生成 4 因素 3 水平 9 次试验的正交实验设计表。

- (1) 操作步骤。
  - ① 打开 SPSS 软件,在选择文件对话框中选择输入数据项,见图 1-2(b),打开一个空数据窗口。
  - ② 按**【数据→正交设计→生成】**顺序单击菜单项,打开**【生成正交设计】**对话框。
  - ③ 在主对话框中定义 4 个因素变量,变量名为 a、b、c、d,变量标签分别为变量名相应的大写字母 A、B、C、D。
  - ④ 逐个选择因素变量,单击**【定义值】**按钮,在相应的对话框中定义因素水平值及其值标签: a [A]: A1、A2、A3; b [B]: B1、B2、B3; c [C]: C1、C2、C3; d [D]: D1、D2、D3。
  - ⑤ 选择**【创建新数据集】**项。在**【数据集名称】**框中输入数据集文件名“ABCD”(默认的扩展名为.sav)将设计结果显示在工作数据文件中 ABCD 中,即当前的数据窗口中。
  - ⑥ 设置随机数种子,选择**【将随机数初始值重置为】**,复选项,在其后框中随便填写一个正整数“2345”。

单击**【选项】**按钮,打开相应对话框。在**【生成的最小个案数】**框中输入“9”。单击**【继续】**按钮返回主对话框。

单击**【粘贴】**按钮,生成如下程序(syntax16-1.sps):

```
SET SEED 2345.
ORTHOPLAN
  /FACTORS=a 'A' (1 'A1' 2 'A2' 3 'A3') b 'B' (1 'B1' 2 'B2' 3 'B3') c 'C'
              (1 'C1' 2 'C2' 3 'C3') d 'D' (1 'D1' 2 'D2' 3 'D3')
  /REPLACE
  /MINIMUM 9.
  _DATASET NAME ABCD.
```

(2) 在工作数据窗口中生成正交设计结果,见图 16-8 (a)。改变随机数种子为 5678。在程序中改变一个语句: SET SEED 5678,其他语句均相同(syntax16-02.sps)。运行结果见图 16-8 (b)。两个设计结果是不同的。

如果在**【数据文件】**栏中选择了**【创建新数据文件】**,并指定了保存位置,则生成的设计



保存在指定位置的数据文件中。程序(syntax16-03.sps)如下：

```
*Generate Orthogonal Design.
SET SEED 2000.
ORTHOPLAN
  /FACTORS=A 'AA' (1 'a1' 2 'a2' 3 'a3') B 'BB' (1 'b1' 2 'b2' 3 'b3') C 'CC'
  (1 'c1' 2 'c2' 3 'c3') D 'DD' (1 'd1' 2 'd2' 3 'd3')
  /OUTFILE='D:\000SPSS 第 5 版\文件存储\正交设计例 1.sav'.
```

如果想要多做两次试验，在【生成正交设计：选项】对话框的【延续个案】栏中选择【延续个案数】，并输入“2”，但是不选择【与其他个案随机混合】。在主对话框中单击【粘贴】按钮，在语句窗中得到如下程序：

```
SET SEED 2000.
ORTHOPLAN
  /FACTORS=a 'A' (1 'A1' 2 'A2' 3 'A3') b 'B' (1 'B1' 2 'B2' 3 'B3') c 'C'
  (1 'C1' 2 'C2' 3 'C3') d 'D' (1 'D1' 2 'D2' 3 'D3')
  /OUTFILE='D:\000SPSS 第 5 版\文件存储\ABCD 保存文件.sav'
  /HOLDOUT 2
  /MIXHOLD NO.
```

运行结果见图 16-9(a)。图 16-9(a)比图 16-8(a)在最后多出两个观测，即第 10、11 行。其 Status\_变量的值标签为“支持”。

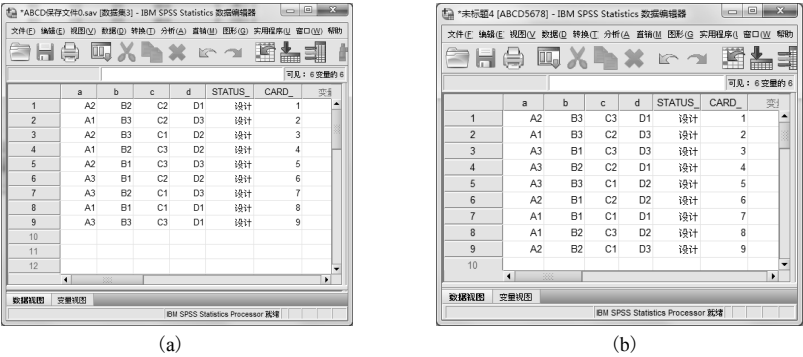


图 16-8 不同随机数种子的 4 因素 3 水平 9 次试验的正交试验设计结果

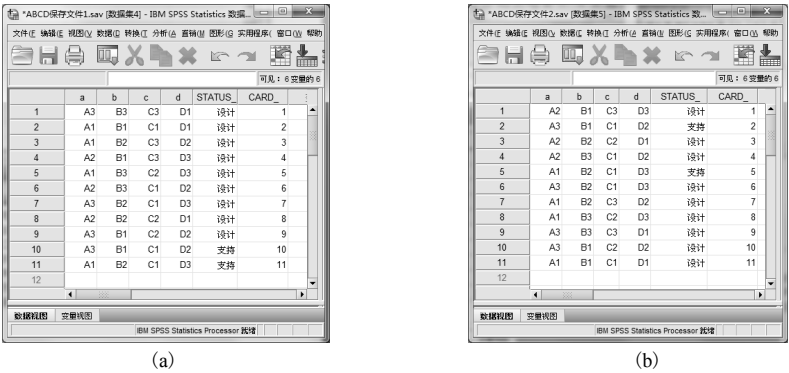


图 16-9 带有延续个案 4 因素 3 水平 9 次试验的正交试验设计结果

如果选择了【延续个案数】，并输入“2”，同时还选择了【与其他个案随机混合】，输出结果随机地将保留观测与试验观测混合，见图 16-9(b)。Status\_变量值标签为支持的观测随机地混在原正交设计的观测中。

## 16.2.5 正交设计过程语句

由于本章的正交试验设计主要为结合分析服务，而窗口式 SPSS 软件的菜单中没有包括结合分析，要想进行结合分析必须编写程序。为编写程序方便，在本章不但介绍结合分析的程序语句，还介绍正交试验设计的程序语句，以便读者一并完成试验设计与分析。

### 1. 正交设计过程 ORTHOPLAN 使用下列语句调用

```
ORTHOPLAN [FACTORS=varlist ['labels'] (values ['labels'])...]
           [{/REPLACE }]    {/OUTFILE='savfile'|'dataset'}    [{/MINIMUM=value}
           [{/HOLDOUT=value} [{/MIXHOLD={YES}}]{NO }
```

其中，“ORTHOPLAN”是命令关键字。ORTHOPLAN 命令为结合分析产生正交的主效应设计。设计结果可以添加在当前的工作数据集中，如果没有工作数据集存在也可以建立一个工作数据集。产生的全组合设计可以列出或使用 PLANCARDS 格式安排。由 ORT-HOPLAN 产生的文件可以用作 CONJOINT 命令要求的设计文件。

### 2. 基本要求

ORTHOPLAN 命令关键字后面跟着 FACTORS 及等号后面的变量列表，如果有变量标签，放在每个变量名后面的引号中。

每个变量名、变量标签后面列出该变量的水平值，如果有值标签，放在值后面的引号中。每个变量的值和值标签列表放在变量名和变量标签后面的括号中。

ORTHOPLAN 在工作数据集中产生观测量。用每个观测量描述结合试验设计的一个侧面，由因素值组合而成。默认是生成最小可能的正交设计。

如果已经存在工作数据集包括正交设计的所有变量，FACTORS 就是可选的子命令。

### 3. 子命令功能与限制

子命令在 ORTHOPLAN 命令语句后面，出现的顺序任意。

(1) 有关的运行结果。

① 如果原来没有工作数据集，ORTHOPLAN 通过 FACTORS 子命令，使用变量和变量值信息建立工作数据集。

② ORTHOPLAN 产生的数据附加在一个工作数据集上。如果没有使用 FACTORS 子命令，因素水平值必须在前面一个 ORTHOPLAN 或 VALUE LABELS 命令定义。

③ 新变量 STATUS\_和 CARD\_如果原来不存在，就产生并附加在 ORTHOPLAN 产生的工作数据集中。试验观测的 STATUS\_变量值为 0，保留观测的 STATUS\_=1，模拟观测的 STATUS\_=2。保留观测由被访者作评价，但是在结合分析 CONJOINT 的效应估计中不用，而是用这些观测作效应估计的合法性检验。模拟观测由用户输入。它们是不由被访者评价的因素水平组合，但是以试验观测的评价为基础由 CONJOINT 进行估计。CARD\_包括在产生的设计中，是观测的标识号。

- ④ 如果试验观测和模拟观测重复, 会给出提示报告。
  - ⑤ 如果用户输入的试验观测 (STATUS\_=0) 是与 ORTHOPLAN 产生的观测相同, 只保留一个。
  - ⑥ 偶尔, ORTHOPLAN 会产生两倍的试验观测。一种处理这些双倍观测的方法是编辑或删除它们。在这些观测中设计不再是正交的。另一种方法是再运行一次 ORTHOPLAN。当设置不同的种子后再运行一次, ORTHOPLAN 可能产生没有重复观测的设计。
  - ⑦ ORTHOPLAN 忽略 SPLIT FILE 和 WEIGHT 命令的作用。
- (2) 限制。
    - ① 不允许有缺失数据。
    - ② 最多可以指定 10 个因素, 每个因素可以指定 9 个水平。
    - ③ ORTHOPLAN 可以产生最多 81 个观测。

#### 4. FACTORS 子命令

FACTORS 子命令指定要在设计中用作因素的变量及其水平值。

(1) 如果数据文件已经存在, 设计产生的观测附加在数据文件上, 是否使用 FACTOR 子命令是可选的; 如果设计产生的观测要保存在建立的新数据集或代替当前已经存在的数据文件, 则必须使用 FACTORS 子命令。

(2) 关键字 FACTORS 后面必须跟变量表, 每个变量的标签是可选的, 每个变量的值列表、值标签是可选的。

(3) 值列表和值标签要用括号, 值可以是数值或是在括号中的字符串。

(4) 可选的变量和值标签要加上省略号。

(5) 如果不用 FACTORS 子命令, 在工作数据集中, 除 STATUS\_ 和 CARD 以外的每个变量都被看成因素变量, 由值标签获得的水平信息在工作数据集定义。ORTHOPLAN 必须在 FACTORS 子命令或 VALUE LABELS 命令中找到变量值信息。

#### 5. REPLACE 子命令

REPLACE 子命令要求用正交设计结果生成或代替当前工作数据集。ORTHOPLAN 可以在数据窗没有数据时运行, 运行结果生成数据占据数据窗。如果数据窗有工作数据集, 此命令用生成的数据集代替当前工作数据集。

(1) 如果使用了 REPLACE, 那么就要求有 FACTORS 子命令。

(2) 默认运行 ORTHOPLAN 的结果不会代替工作数据集。在 FACTORS 子命令中指定的新变量加上变量 STATUS\_ 和 CARD\_ 附加在工作数据集上。

(3) 当前工作数据集中的数据对要建立的设计文件来说没有作用时, 需要使用 REPLACE。工作数据集将被有 STATUS\_、CARD\_ 变量的和任何其他在 FACTORS 子命令中指定的变量所代替。

#### 6. OUTFILE 子命令

OUTFILE 子命令把正交设计结果保存到 SPSS 数据文件。

对输出文件只需指定文件名, 可以是文件名或以前宣告过的数据集的名字。

(1) 默认不创建新数据文件。任何用 FACTORS 指定的新变量加上 STATUS\_ 和 CARD\_ 附加在工作数据文件中。

- (2) 输出数据文件包括 STATUS\_、CARD\_和所有 FACTORS 子命令中指定的变量。
- (3) 由 OUTFILE 产生的文件可以用于其他命令语句，如 PLANCARDS 和 CONJOINT。
- (4) 如果使用了 OUTFILE，可以不用 REPLACE。

7. MINIMUM 子命令

MINIMUM 子命令指定最小观测数。

- (1) 不用此命令，默认产生正交设计必需的最小观测数。
- (2) MINIMUM 后面跟着正整数，这个正整数要小于或等于所有可能的水平组合所能形成的观测总数。
- (3) 如果 ORTHOPLAN 不能产生 MINIMUM 所要求的至少的观测数，就产生适合指定的因素数和水平数的最大数。

8. HOLDOUT 子命令

HOLDOUT 子命令按关键字后面的数字产生附加在正规设计上的保留观测。保留观测由被访者评价，但是在 CONJOINT 估计效应时不使用它。

- (1) 不指定 HOLDOUT 就不产生保留观测。
- (2) HOLDOUT 后跟正整数，这个正整数要小于等于由所有可能的因素水平组合所形成的观测总数。
- (3) 保留观测由另一个随机设计产生，不是主效应试验设计。保留观测不会复制试验观测，也不会彼此复制。
- (4) 试验观测和保留观测是在生成的设计中随机混合在一起还是保留观测附加在试验观测后面，取决于 MIXHOLD 子命令。保留观测的 STATUS\_变量值是 1。任何模拟观测都安排在试验观测和保留观测后面。

9. MIXHOLD 子命令

MIXHOLD 子命令指定保留观测是随机地与试验观测混合还是单独出现在文件的试验设计后面。如果没有指定 MIXHOLD，默认是“NO”，即在文件中保留观测将出现在试验观测的后面。

- (1) MIXHOLD 后面跟着关键字“YES”，要求把保留观测与试验观测随机混合。
- (2) 没有 HOLDOUT 子命令，指定 MIXHOLD 无效。

10. 程序举例

【例 2】 酸奶的市场调查试验设计程序如下（见 YUGPLAN1.SPS）：

```
ORTHOPLAN
/FACTORS=weight '重量' (600 '600g' 800 '800g' 1000 '1kg') Warranty
'保质期' (3 '3天' 5 '5天' 7 '7天') Casing '包装' (1 '纸盒' 2 '瓶子')
/REPLACE.
__DATASET NAME YUGPLAN1.
```


- (1) 程序解释。

① ORTHOPLAN 命令后面的“FACTOR”子命令定义了 3 个变量及其水平值和值标签：weight 变量，标签为“重量”，其值有 3 个水平，以 g 作单位，600、800、1000 三个水平值标

签分别为“600g”、“800g”、“1000g”；Warranty 变量，标签为“保质期”，有 3 个水平值 3、5、7 标签，分别为“3 天”、“5 天”、“7 天”；Casing 变量，标签为“包装”，有 2 个水平值 1、2，分别表示纸盒包装和塑料包装。这些变量及其水平将被用于生成试验设计文件。

② REPLACE 命令要求设计结果放在一个数据集中，该数据集将是打开的数据文件，在数据窗口中可以看到结果。SPSS 的某些版本会把设计结果代替原打开的数据集，所以新建一个空数据集比较稳妥。SPSS 20.0 版的该语句会另建一个新数据集，标题栏中除用 DATANAME 语句定义的文件名外，在该文件名前面还有“未标题”字样，待真正保存了才标有赋予它的文件名。

③ DATASET NAME 语句设置该生成设计结果的数据集名称为“YUGPLAN1.sav”。是在新数据窗口中生成的，还需要保存成外部文件，否则结束 SPSS 运行后，将丢失设计结果。

(2) 在语句窗口中，用鼠标选择程序语句，单击运行图标按钮，提交系统执行。执行结果见图 16-10。

**【例 3】** 在语句窗口输入以下程序(YUGPLAN2.sps)：


```
ORTHOPLAN FACTORS=weight '重量' (600 '600g' 800 '800 g' 1000 '1kg')  
WARRANTY '保质期' (3 '3天' 5 '5天' 7 '7天') casing '包装' ( 1 '纸盒' 2 '瓶子' )  
/MINIMUM=9 /HOLDOUT=6.  
/REPLACE.  
DATASET NAME YUGPLAN2.
```

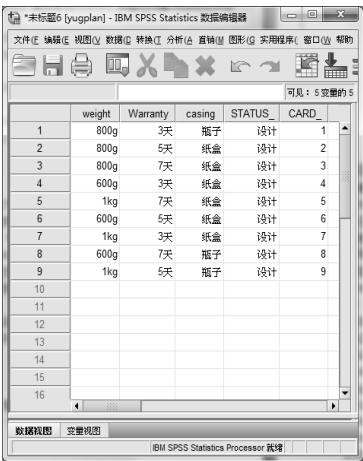
(1) 程序解释。

① ORTHOPLAN 命令后面的 FACTOR 子命令定义了 3 个变量及其水平值和值标签，与【例 2】一样，这些变量及其水平将被用于生成试验设计文件。不同的是/MINIMUM=9/HOLDOUT=6。

② MINIMUM 子命令指定正交试验设计至少要生成 9 个试验观测；HOLDOUT 子命令指定要生成 6 个保留观测。

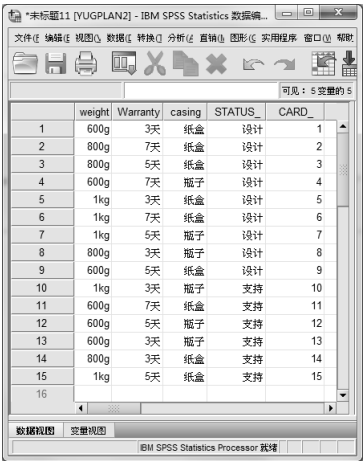
③ DATASET NAME 语句设置该生成设计结果的数据集名称为“YUGPLAN2.sav”。

(2) 在语句窗口中，用鼠标选择程序语句，单击运行图标按钮，提交系统执行。执行结果见图 16-11。



	weight	Warranty	casing	STATUS	CARD	
1	800g	3天	瓶子	设计	1	
2	800g	5天	纸盒	设计	2	
3	800g	7天	纸盒	设计	3	
4	600g	3天	纸盒	设计	4	
5	1kg	7天	纸盒	设计	5	
6	600g	5天	纸盒	设计	6	
7	1kg	3天	纸盒	设计	7	
8	600g	7天	瓶子	设计	8	
9	1kg	5天	瓶子	设计	9	
10						
11						
12						
13						
14						
15						
16						

图 16-10 【例 2】程序运行结果



	weight	Warranty	casing	STATUS	CARD	
1	600g	3天	纸盒	设计	1	
2	800g	7天	纸盒	设计	2	
3	800g	5天	纸盒	设计	3	
4	600g	7天	瓶子	设计	4	
5	1kg	3天	纸盒	设计	5	
6	1kg	7天	纸盒	设计	6	
7	1kg	5天	瓶子	设计	7	
8	800g	3天	瓶子	设计	8	
9	600g	5天	纸盒	设计	9	
10	1kg	3天	瓶子	支持	10	
11	600g	7天	纸盒	支持	11	
12	600g	5天	瓶子	支持	12	
13	600g	3天	瓶子	支持	13	
14	800g	3天	纸盒	支持	14	
15	1kg	5天	纸盒	支持	15	
16						

图 16-11 【例 3】程序运行结果

注意：这个程序可以使用中文变量标签和值标签。

数据编辑窗是空窗口，无任何数据时，将产生试验设计数据，包括 5 个变量：WEIGHT、WARRANTY、CASING、STATUS\_和 CARD\_。

数据包括 15 个观测，其中试验观测 9 个，其 STATUS\_变量值为 0；还有 6 个保留观测，其 STATUS\_值为 1，排列在 9 个试验观测后面，见数据文件 data16-11。

应该说明的是，该程序生成的观测置于数据窗中，形成当前工作数据集。

如果该程序增加一个子命令 OUTFILE='YUGPLAN.sav'.则将生成的正交设计保存在命名为“YUGPLAN.sav”的数据文件中。

输出还有警告信息：系统已经成功生成含有 9 的计划。

如果用户加入的保留观测比较多，有时会破坏正交性，系统也会在警告信息中指出。

**【例 4】** 带有模拟侧面观测生成的程序。命令语句见文件 YUGPLAN2.sps，内容如下：

```
DATA LIST FREE /WEIGHT  WARRANTY CASING.
VARIABLE LABELS WEIGHT '重量'  WARRANTY "保质期" CASING "包装".
VALUE LABELS weight 600 '600g ' 800 '800 g' 1000 '1kg'
/warranty 3 '3 天' 5 '5 天' 7 '7 天'
/casing 1 '纸盒' 2 '瓶子'.
BEGIN DATA
1000 5 1
1000 3 2
END DATA.
ORTHOPLAN.
```

(1) 程序解释。

① DATA LIST 命令语句定义了 3 个自由格式的变量 WEIGHT、WARRANTY、CASING。

② VARIABLE LABELS 语句定义了 3 个变量的变量标签。

语句规则是：关键字 VARIABLE LABELS 后面的每个变量名后面引号(单、双引号均可)中的是它的变量标签，最后用圆点结束该语句。

③ VALUE LABELS 命令语句定义了 3 个变量的水平值和值标签：WEIGHT 变量 3 个水平，值为 600、800、1000，标签表明单位为“g”，是酸奶的不同重量；WARRANTY 变量 3 个水平，值为 3、5、7，表明保质期的 3 个水平为 3 天、5 天、7 天；CASING 变量 2 个水平，值为 1、2，表示两种包装：纸盒、瓶子。

语句规则是：关键字 VALUE LABELS 后每个变量后面的是它的水平值，每个值后面的引号中是值标签，用圆点结束该语句。

**注意：**这里的解释是中文，而程序中变量名是用英文。变量标签和值标签均可以使用中文，如上述程序中的值标签。

④ BEGIN DATA –END DATA 语句中，两行数字是按 DATA LIST 语句中的变量顺序，给出 2 个观测的 3 个变量值。这两个观测在正交试验设计中作为模拟侧面观测生成的来源。

⑤ ORTHOPLAN 语句使用上述数据作为正交试验设计的因素、水平和模拟侧面的观测，无须再使用 FACTORS 子命令定义试验设计需要使用的变量及其水平值。

同样，如果工作数据文件中已经存在设计需要的变量、值及值标签，需要的模拟观测也已经输入，那么 ORTHOPLAN 语句无须任何子命令就把数据窗口中的所有变量当作设计需要的因素。

(2) 生成的设计结果见数据文件 data16-02。除程序中定义的变量外，还有两个变量，变量

CARDS\_值为观测号，每类观测自行编号；变量 STATUS\_值表明观测的性质。有 9 个观测的 STATUS\_值为 0，它们是试验观测；2 个 STATUS\_值为 2 是模拟观测；共 11 个观测。

变量有 3 个，它们的水平数分别是 3、3、2，水平的全组合数是 18，输出默认的 9 种试验组合，即酸奶属性的 9 个侧面。

注意：如果原数据窗中的数据量很大，变量名与要生成的设计中的变量名还相同，则最好将数据窗中的数据清除后再运行新程序，或者通过单击菜单项【文件→新建→数据】建立一个空的数据文件，再进行正交设计的各步骤。



	WEIGHT	WARRANTY	CASING	STATUS_	CARD_
1	600g	5 天	纸盒	设计	1
2	600g	3 天	纸盒	设计	2
3	1kg	5 天	纸盒	设计	3
4	800 g	5 天	瓶子	设计	4
5	800 g	3 天	纸盒	设计	5
6	600g	7 天	瓶子	设计	6
7	1kg	3 天	瓶子	设计	7
8	800 g	7 天	纸盒	设计	8
9	1kg	7 天	纸盒	设计	9
10	1kg	5 天	纸盒	模拟	1
11	1kg	3 天	瓶子	模拟	2
12					

图 16-12 【例 4】程序运行结果

### 16.3 试验设计结果的打印

在结合分析的研究中，试验设计结果要在进行调查时显示给被访者，请被访者评分或排序。SPSS 的正交设计的显示功能可以两种方式显示或打印正交设计结果：

- ① 以粗略的列表格式打印设计结果，以便撰写报告或存档。
- ② 可以显示给被访者观看的，产品的每个属性组合一个个列出的格式。

结果打印程序还允许用户自己加上标题，每个标题占一行；可以打印空行；允许用户加注脚，每个注脚占一行，也可以打印空行。

如果选择列表格式打印，在列表之前打印标题，列表最后打印注脚。

如果选择一个一个列出的格式打印，还可以在每个属性组合之前打印标题，每个属性组合之后打印注脚。

#### 16.3.1 设计结果打印过程

从主菜单中顺序单击【数据→正交设计→显示】，打开正交设计结果【显示设计】主对话框，见图 16-13。操作步骤如下：

- (1) 在左侧的变量表中选择正交设计的全部因素，将其移到右侧的【因子】框中。
- (2) 在【格式】栏中选择打印方式。
  - ①【试验者列表】。对试验侧面使用列表方式显示或打印；分别打印试验侧面、保留侧面，并在它们后面列出模拟侧面。
  - ②【群体配置文件】。打印要呈现在被访者面前的设计侧面，不区分保留侧面、模拟侧面。运行结果是打印全部卡片。
  - (3) 设置标题和注脚。单击【标题】按钮，打开如图 16-14 所示的【显示设计：标题】对话框。
    - ① 在【配置文件标题】框内输入标题，也可以输入对被访者的提示，例如排序须知或者打分方法等说明文字等。
    - ② 在【配置文件页脚】框内输入页脚。

单击【继续】按钮返回主对话框。

在主对话框中单击【确定】按钮提交运行。



图 16-13 【显示设计】主对话框



图 16-14 【显示设计：标题】对话框

16.3.2 打印调查用卡片实例

【例 5】以酸奶的偏好调查为例。在 16.3.1 节中做好的正交试验设计保存在数据文件 data16-01 中，包括 9 个实验侧面、6 个支持(保留)侧面。

(1) 打开文件后的操作如下：

① 在主菜单中顺序单击【数据→正交设计→显示】，打开正交设计结果【显示设计】主对话框，见图 16-13。

② 在左侧的变量表中选择 weight、warratny、casing 三个因素变量，将其移到右边的【因子】栏中。

③ 在【格式】栏中选择两种打印方式：以列表方式和卡片方式输出。【试验者列表】方式为自己保留设计结果。【群体配置文件】方式实际是卡片方式，即各侧面一一分别列出，需要呈现在被访者面前时使用。

④ 单击【标题】按钮，打开【显示设计：标题】对话框，输入标题和页脚。

● 在【配置文件：标题】框中：

第 1 行空出，回车后，从第 2 行开始输入自己打印的标题。因为第 1 行会有系统默认的标题出现。输入的标题是“《酸奶偏好调查》”。

第 3 行输入给被访者的提示“请先查看所有酸奶卡片，按你喜好的顺序排列”。

第 4 行输入“最喜欢的在卡片上标 1，次之标 2，依此类推”。

第 5 行为空行。

● 在【配置文件：页脚】栏中输入感谢语“谢谢您的参与!”。

⑤ 单击【继续】按钮返回主对话框。在主对话框中单击【确定】按钮提交运行。

(2) 执行的程序如下：

```
PLANCARDS
  /FACTOR=WEIGHT WARRANTY CASING
  /FORMAT BOTH
  /TITLE '《酸奶偏好调查》请先查看所有酸奶卡片，按你喜好的顺序排列。最喜欢的在卡片上
        标 1，次之标 2，依此类推。'
  /FOOTER '                                     谢谢您的参与!'
```

不难读懂这个程序：PLANCARDS 是命令关键字；FACTORS 子命令定义了 3 个因素变量；FORMAT 子命令选择了两种打印方式；TITLE 子命令在引号中给出标题字符串；在 FOOTER 子命令给出的是作为注脚的字符串。



(3) 输出结果见图 16-15 和图 16-16。

图 16-15 是列表式输出，试验侧面和保留侧面都显示在一个表中。

图 16-16 没有列出全部卡片，只列出给一个被调查者的第一个试验侧面，见图 16-16 上半部；下半部是保留侧面，即第 15 个卡片。

《酸奶偏好调查》  
请先查看所有酸奶卡片，按你洗好的顺序排列  
最喜欢的在卡片上标**1**，次之标**2**，依此类推。

	卡标识	重量	保质期	包装
1	1	800 g	3天	瓶子
2	2	600g	3天	纸盒
3	3	800 g	5 天	纸盒
4	4	600g	5 天	纸盒
5	5	800 g	7天	纸盒
6	6	1kg	3天	纸盒
7	7	600g	7天	瓶子
8	8	1kg	5 天	瓶子
9	9	1kg	7天	纸盒
10 <sup>a</sup>	10	600g	5 天	瓶子
11 <sup>a</sup>	11	800 g	5 天	瓶子
12 <sup>a</sup>	12	1kg	7天	瓶子
13 <sup>a</sup>	13	600g	7天	纸盒
14 <sup>a</sup>	14	600g	3天	瓶子
15 <sup>a</sup>	15	1kg	3天	瓶子

感谢您的参与！  
a. 保留

图 16-15 各侧面列表式输出

概要文件编号 **1:**  
《酸奶偏好调查》  
请先查看所有酸奶卡片，按你洗好的顺序排列  
最喜欢的在卡片上标**1**，次之标**2**，依此类推。

卡标识	重量	保质期	包装
1	800 g	3天	瓶子

感谢您的参与！

概要文件编号 **15:**  
《酸奶偏好调查》  
请先查看所有酸奶卡片，按你洗好的顺序排列  
最喜欢的在卡片上标**1**，次之标**2**，依此类推。

卡标识	重量	保质期	包装
15	1kg	3天	瓶子

感谢您的参与！

图 16-16 调查用卡片式输出

16.3.3 正交试验设计打印过程语句

```
PLANCARDS [FACTORS=varlist]
[/FORMAT={LIST}] {CARD} {BOTH}
[/TITLE='string'] [/FOOTER='string'] [/OUTFILE=file]
```

PLANCARDS 为结合分析产生供保留的侧面清单或卡片。设计文件由 ORTHOPLAN 生成或用户输入。打印的侧面可以用于被访者评价偏爱项目的试验依据。

此命令读取工作数据集中的数据，以 FACTORS 子命令定义的因素变量为打印的变量，按后续语句要求打印正交试验设计的结果。

过去版本的 PAGINATE 子命令已经作废。不能在这里使用。

(1) 几点说明。除了 FACTORS 子命令外都是可选的子命令。基本命令就是 PALNCARDS。不指定 FACTORS，该命令使用当前工作数据集中除 STATUS\_和 CARD\_以外的所有变量作为打印的基本变量。

① PLANCARD 假定工作数据集是结合研究的正交设计结果。在这样的文件中，每个观测就是一个结合试验设计的侧面。

② PLANCARD 使用在工作数据文件中由 ORTHOPLAN 产生或由 VARIABLE 和 VALUE LABELS 命令产生的因素和因素水平标签。

③ **SPLIT FILE** 命令对单个卡片输出方式无效。在列表的格式中,每个侧面描述一个不同的设计,并且该命令对每个 **SPLIT FILE** 产生的子文件开始一个新的列表。

④ **WEIGHT** 命令对 **PLANCARD** 命令无效。

⑤ 不把缺失值当作缺失值,而当作一个有效值看待。

(2) **FACTORS** 子命令识别要用作因素的变量和它们的标签出现在输出中的顺序。该子命令允许定义字符串变量。

① 关键字 **FACTORS** 后面跟变量表。

② 如果没指定 **FACTORS**,默认工作数据集中除 **STATUS\_**和 **CARD\_**外的所有变量,以它们在文件中出现的顺序作为因素使用。

(3) **FORMAT** 子命令指定一个如何显示侧面的方式。选项是列表方式(关键字 **LIST**)和单个侧面方式(关键字 **CARD**)。

① 关键字 **FORMAT** 后面跟着 **LIST**、**CARD** 或 **BOTH**(**ALL** 也可代替 **BOTH**)。

② 默认的格式是 **LIST**。

③ 用 **LIST** 格式,以试验侧面、保留侧面、模拟侧面的顺序列出。用 **CARD** 格式,保留侧面作为一个卡片输出,不产生模拟侧面输出。

如果 **FORMAT=LIST** 与 **OUTFILE** 子命令一起被指定,则 **OUTFILE** 子命令无效。**OUTFILE** 只对 **CARD** 格式有效。与 **FORMAT=BOTH** 一起指定 **OUTFILE** 时,与 **OUTFILE**、**FORMAT=CARD** 一起使用的效果是相等的。

(4) **OUTFILE** 子命令命名一个外部文件,以单个侧面格式写入。列表方式不写到这个外部文件中。

① 默认没有到外部文件的输出。

② **OUTFILE** 关键字后面跟着一个外部文件,该文件以系统通常的形式指定。

③ 如果 **OUTFILE** 子命令用 **FORMAT=LIST** 一起指定,则 **OUTFILE** 子命令无效。**OUTFILE** 子命令仅施加于 **FORMAT=CARD**。

(5) **TITLE** 子命令指定用于输出的标题,无论是列表格式还是单一侧面格式的顶部的标题字符串。

① 提供默认的标题,除非用 **OUTFILE** 子命令直接输出到一个外部文件。

② 关键字 **TITLE** 后面跟着用单引号引起来的字符串。

③ 如果标题中有单引号,可以用双引号代替单引号把字符串引起来。

④ 每个 **TITLE** 子命令可以指定多个字符串;每个字符串将出现在不同的行中。

⑤ 使用空字符串('')会出现一个空行。

⑥ 可以使用多个 **TITLE** 子命令;每个子命令出现在单独的行上。

(6) **FOOTER** 子命令指定的字符串将出现在列表格式输出的底部,或者单独侧面格式输出的底部。

① 如果 **FOOTER** 在程序中没有使用,则列表和卡片底部是空的。

② **FOOTER** 后面跟着一个放在单引号中的字符串。

③ 每个 **FOOTER** 子命令可以指定多个字符串。每个字符串出现在一个单独的行上。

④ 使用空字符串产生空行。

⑤ 可以指定多个 **FOOTER** 子命令,每个子命令会出现在一个单独的行上。

## 16.4 结合分析的语句与编程

打印了供调查使用的卡片, 经过培训的调查员就可以按抽样设计矩形调查了。得来的数据经过整理, 需要进行结合分析, 结合分析的结果提供给决策者作为决策依据。

在 SPSS 窗口式运行方式的分析菜单中没有结合分析, 要进行结合分析必须进行编程。下面介绍结合分析命令语句及编程要领。

### 16.4.1 结合分析过程语句

#### 1. 结合分析过程 CONJOINT 使用下列命令语句调用

```
CONJOINT [PLAN={* }]{'savfile'|'dataset'}
[/DATA={* }]{'savfile'|'dataset'}/{SEQUENCE}=varlist {RANK }{SCORE }
[/SUBJECT=variable]
[/FACTORS=varlist['labels'] [{DISCRETE[{MORE}]]]{ {LESS} }{LINEAR[{MORE}]] }
{ {LESS} }{IDEAL }{ANTIIDEAL }[values['labels']]varlist...
[/PRINT={ALL** } [SUMMARYONLY]]{ANALYSIS }{SIMULATION }{NONE }
[/UTILITY=file]
[/PLOT={ [SUMMARY] [SUBJECT] [ALL] }]{ [NONE**] }
```

CONJOINT 分析偏爱分数或秩数据。由 ORTHOPLAN 产生的或由用户输入的设计文件描述了在偏爱项目的研究中被打分或排秩的全概念数据集。一种连续或离散模型用于估计每个被访者的或一组被访者的效应。

可以指定怎样把被期望的因素与分数或秩联系起来。

输出可以包括试验数据或模拟数据的分析, 或两者都包括。

对每个被访者的效应估计和有关的统计量可以输出到 SPSS 外部数据文件, 以供进一步分析或作图。

#### 2. 基本规范

以下基本规范会涉及一些语句, 在讲述有关语句的要求和使用方法时不再重复。

(1) 要求有 CONJOINT、PLAN 或 DATA 子命令和 SEQUENCE、RANK 或 SCORE 子命令描述数据类型。

(2) CONJOINT 要求必须有两个文件: 设计文件和数据文件。PLAN 子命令指定设计文件 DATA 子命令指定调查数据文件。不一定同时使用两个子命令, 可以用 PLAN 或 DATA 子命令指定一个文件, 而当前的工作数据集作为另一个文件。

(3) 默认使用 DISCRETE 模型对设计文件中的所有变量(除了名为 STATUS\_ and CARD 的变量)计算估计效应。输出包括 Kendall's tau 和皮尔逊积矩相关系数, 预测分数和实际分数之间的相关, 显示单尾检验的显著性水平。

(4) 子命令可以是任意顺序的。

(5) 可以执行多个 FACTORS 子命令, 而其他子命令, 如果一个程序中出现多个, 只有最后一个可以执行。

(6) 设计文件和数据文件都可以是外部 SPSS 数据文件。

在设计文件中, 试验侧面的变量 STATUS\_ 必须为 0, 保留侧面 STATUS\_ 必须为 1, 模拟

侧面的 STATUS\_ 值必须为 2。保留侧面由被访者评价,但 CONJOINT 估计效应时不用,而是在检验效应合法性时使用。模拟侧面是没有被被访者评价的,但因素水平组合由 CONJOINT 根据试验侧面的评价估计其效应。如果没有 STATUS\_ 变量,设计文件中的所有侧面都被假定为试验侧面。

(7) 设计文件中的所有变量除了 STATUS\_ 和 CARD\_ 都被 CONJOINT 作为因素使用。

(8) 除了对每个被访者进行估计外,对每个在数据文件中定义的分开的数据组计算平均效应。

(9) CONJOINT 检验因素的正交性。如果所有因素都不正交,显示 Cramér 的  $V$  矩阵统计量,描述非正交性。

(10) 在使用 SEQUENCE 或 RANK 数据时,CONJOINT 对秩尺度进行中心转换,以使计算的系数为正。

(11) 设计文件在收集数据以后不能排序或以任何方法修改,因为设计文件中的侧面顺序必须与数据文件中的数值顺序一一对应(CONJOINT 使用的侧面顺序要与它们在设计文件中出现的顺序一致),不是 CARD\_ 的值决定侧面顺序。如果数据记录方法是 RANK 或 SCORE,数据文件中第一个被访者的第一个回答就是设计文件中第一个侧面的秩或分数;如果数据记录方法是 SEQUENCE,数据文件中第一个被访者的第一个回答是最偏爱的侧面的侧面号(由设计文件中的侧面顺序决定)。

### 3. 限制

(1) 因素必须是数值型变量。

(2) 设计文件不能包括缺失值或观测的权重。在工作数据集中, SUBJECT 变量带有缺失值的侧面被聚在一起并在最后计算平均值。如果有被访者的任何一个偏爱数据(秩、分数或侧面号)是缺失的,那个被访者的数据就被跳过,不参与分析。

(3) 因素必须至少有 2 个水平,每个因素水平最大值为 99。

### 4. PLAN 子命令识别包括全概念侧面的文件

(1) PLAN 子命令关键字后面跟着引用的 SPSS 数据文件名或当前打开的包括设计的数据集的文件说明。星号代表工作数据集是设计文件。

(2) 程序中如果没有 PLAN 子命令,当前工作数据文件被认为是默认的设计文件。工作数据文件不能再用 DATA 或 PLAN 子命令指定为设计文件或数据文件。

(3) 设计文件可以由 ORTHOPLAN 产生,也可由用户直接输入。设计文件可以包括 CARD\_ 和 STATUS\_ 变量,并且必须包括结合分析研究的因素。

### 5. DATA 子命令指定包括被访者的偏爱分数或秩的(调查数据)文件

(1) DATA 子命令关键字后面跟着被指定的 SPSS 数据文件或用星号指定当前在数据窗打开的包括调查数据的数据集文件。

(2) 如果程序中没有 DATA 子命令,当前工作数据集就是默认的调查数据文件。工作数据文件不能再用 DATA 或 PLAN 子命令指定为设计文件或数据文件。

(3) 在数据文件中,一个变量可以是被访者的标识变量,所有其他变量是被访者的回答数据,并在数量上等于设计文件中的实验侧面和保留侧面的总数。

(4) 被访者的回答可以以秩的形式赋予安排好的侧面顺序,或以分数赋予安排好的侧面顺序,或者侧面号按从最喜欢到最不喜欢的顺序安排。

(5) 允许秩或分数存在结(秩或分数相同的观测)。如果出现了秩结, CONJOINT 发布警告信息然后继续分析。数据以 SEQUENCE 顺序格式记录时不能有结, 因为每个侧面号必须是唯一的。

## 6. SEQUENCE、RANK、SCORE 子命令指定偏爱数据记录的方法

(1) 必须从三个子命令中选择指定一个, 而且只有一个。

(2) 每个子命令后面列出包含偏爱数据的变量名(侧面号变量、秩或分数变量)在设计文件中有多少试验侧面和保留侧面, 就必须列出多少变量名。

(3) 子命令关键字含义与规定。

① SEQUENCE 数据文件中的每个数据点是一个侧面号, 以最偏爱的侧面开始并以最不偏爱的侧面结束, 如被访者被问及从最喜欢到最不喜欢排列侧面卡片。研究人员记录哪个侧面号是第一个, 哪个侧面号是第二个, 等等。

② RANK 每个数据点是秩, 从侧面 1 的秩开始, 然后是侧面 2 的秩, 依此类推。这就是被访者被要求对每个侧面安排一个秩(顺序), 秩从 1 到  $n$ , 这里的  $n$  是侧面数。较低的秩意味较高的偏爱。

③ SCORE 每个数据点是赋予该侧面的偏爱分数, 以侧面 1 分数开始, 然后是侧面 2 分数, 依此类推。例如, 通过要求被访者给出 1~100 的值, 表明他们有多喜欢这个侧面, 高分数对应高偏爱, 就是这样的数据类型。

## 7. SUBJECT 子命令指定一个标识变量

所有这个变量具有相同值的观测被组合以便估计效应。

(1) 如果没有使用 SUBJECT, 所有数据都被假设来自一个被访者, 输出并仅显示一组摘要。

(2) SUBJECT 后面跟着变量名。这个变量值标识被访者, 或者标识一组被访对象。

## 8. FACTORS 子命令指定要分析的每个因素与秩或分数相关的类型

(1) 如果没有使用 FACTOR 子命令, 则对所有因素假设为离散模型。

(2) 在设计文件中的所有变量, 除了 CARD\_ 和 STATUS\_ 外都被用作因素, 即使它们没有 FACTOR 子命令中出现。

(3) FACTOR 后面跟着变量列表、模型和括号中的模型说明, 该说明描述在秩或分数与变量列表的因素水平之间的期望关系。

(4) 模型说明由模型名和与指定模型的选项组成, 对 DISCRETE、LINEAR 模型的选项 MORE 或 LESS 关键字表明所期望关系的趋势, 还可以指定值和值标签。

(5) MORE 和 LESS 关键字对估计效应不起作用。它们被简单地用作识别那些估计与期望的趋势不一致的观测量(被访者)。4 个可用的模型如下:

① DISCRETE 因素水平是分类的, 不作因素和分数或秩之间关系的假设。这个设置是默认的。在 DISCRETE 后面指定关键字 MORE 表明因素的较高水平被期望是更偏爱; 指定关键字 LESS 表明因素的较低水平被期望是更偏爱。

② LINEAR 线性关系。假设期望分数或秩与因素水平的关系是线性的。在 LINEAR 后面指定关键字 MORE 表明因素的较高水平被期望是更偏爱; 指定关键字 LESS 表明因素的较低水平期望是更偏爱。

③ IDEAL 二次关系描述渐减的偏爱。假设期望分数或秩与因素水平之间是二次关系。它假设存在一个理想的因素水平。与这个理想点的距离,在任意一个方向上都是与渐减的偏爱相联系。用这个模型描述的因素应该至少有 3 个水平。

④ ANTI IDEAL 二次关系描述渐增的偏爱。假设期望分数或秩与因素水平之间是二次关系。它假设存在一个最差的因素水平。与这个最差点的距离,在任意一个方向上都是与渐增的偏爱相联系。用这个模型描述的因素应该至少有 3 个水平。

(6) 对那些没有列在 FACTOR 子命令中的变量,都假设 DISCRETE 离散模型,即无模型假设。

(7) 当 MORE 或 LESS 关键字与 DISCRETE 或 LINEAR 一起使用时,如果所期望的趋势不出现(发生),则会给出注释。

(8) IDEAL 和 ANTIIDEAL 两者都生成因素的二次方程,唯一的差别是从与特定点出发,偏爱增加还是减少。对这两个模型的效应估计都相同。当所期望的模型不存在时,会给出提示。

(9) 选择的值和值标签列表允许记录数据和(或)修改值标签。新值以它们出现在值列表中的顺序以最小的现有值开始替换已经存在的值。如果新值没有指定给一个已经存在的值,则该值保持不变。

(10) 新值标签在引号中指定。没有新标签的新值保持现有的标签;新值标签按照它们出现的顺序赋予新值,如果没有新值赋予它,以最小的存在的值开始。

(11) 对每个记录的因素显示一个标签,显示原始的记录值和值标签。

(12) 如果因素水平是离散的分类代码(如 1、2、3),这些值就是 CONJOINT 在计算中使用的值,即使值标签包含实际值(如 600、800、1000),但值标签不会用于计算。用户可以用如上所述的值重新编码,改变代码为实际值。重新编码不会影响 DISCRETE 因素但是改变了 LINEAR、IDEAL 和 ANTIIDEAL 因素的系数。

(13) 对变量的描述输出顺序是所有的 DISCRETE 变量、LINEAR 变量、IDEAL 和 ANTIIDEAL 因素出现在 FACTOR 子命令中的顺序。

## 9. PRINT 子命令控制输出的内容

PRINT 子命令控制输出是否包括对试验数据、对模拟数据的分析结果,或两种都包括,或没有输出。下列关键字可以使用:

(1) ANALYSIS 输出仅包括试验数据分析的结果。

(2) SIMULATION 仅输出模拟数据的分析结果。3 个模拟模型是:最大效应模型、Bradley-Terry-Luce 即 BLT 模型和对数模型。

(3) SUMMARYONLY 输出仅包括综合性的概述。如果被访者很多,就可以看到综合概述,没必要对每一个被访者都有输出,使输出量很大。

(4) ALL 输出包括试验数据和模拟数据的分析。ALL 是默认的。

(5) NONE 没有输出内容。当仅想把分析结果写到效应文件时,使用该关键字。

## 10. UTILITY 子命令把效应分析结果写到指定的 SPSS 数据文件中

(1) 程序中没有 UTILITY 子命令,就没有效应文件输出。

(2) UTILITY 后面跟着要输出的效应文件名。

(3) 该文件使用操作系统惯用的指定方法指定。

(4) 效应文件对每个被访者有一个观测。如果没有使用 SUBJECT,效应文件包括一个单独的观测的统计量,把这组观测作为一个整体。

(5) 写到效应文件中的变量按下列顺序安排：

- ① 任意一个工作数据集中的 SPLIT FILE 变量。
- ② 任意一个 SUBJECT 变量。
- ③ 回归方程的常数项。对应的变量名为 CONSTANT。

④ 对 DISCRETE 因素，对被访者估计所有效应。对所有 DISCRETE 因素估计的效应变量名为因素名后面跟着数字构成。第一个效应后面是 1，第二个效应后面为 2，依此类推。

⑤ 对 LINEAR 因素，给出单个系数。LINEAR 因素的效应名是因素名后面跟着\_L(预测分数的计算是因素值乘系数)。

⑥ 对 IDEAL 或 ANTIIDEAL 因素，因为是二次模型，所以给出两个系数。系数变量名的命名是在因素名后面分别加\_L(一次项系数)和\_Q(二次项系数)构成的(要使用这些系数计算预测分数，需用因素值乘以第一个系数加上第二个系数与因素值的平方的乘积)。

⑦ 对设计文件中的所有侧面估计秩或分数。估计的秩或分数的名字对试验和保留侧面用 SCORE<sub>n</sub>，对模拟侧面用 SIMUL<sub>n</sub>。这里的 *n* 是在设计文件中的位置顺序。即使数据是秩，试验和保留侧面的名字也是 SCORE。

如果生成的变量名太长，在添加新后缀之前，从原始变量名末尾截掉字母。

11. PLOT 子命令

除了 CONJOINT 产生的输出外还生成图形。下面是可以用作子命令参数的关键字。

(1) SUMMARY：对所有变量产生重要性价值条形图和每个变量的效应条形图。如果使用 PLOT 子命令时没有给出关键字，该关键字是默认的。

(2) SUBJECT：对每个因素的重要性价值绘制一簇条形图，由被访者构成簇，每个因素一簇条形图，表明每个因素水平、每个被访者的效应。如果没有 SUBJECT 子命令指定变量的命名，就不产生图形，并显示警告信息。

(3) ALL：绘制 SUMMARY 和 SUBJECT 两种图。

(4) NONE：不产生任何图形。如果该子命令被省略，该设置是默认的。

16.4.2 结合分析语句实例

【例 6】 一个结合分析程序的基本结构。

```
CONJOINT PLAN='/DATA/CARPLAN.SAV'                                ①
/FACTORS=WEIGHT (LINEAR MORE) WARRANTY (DISCRETE MORE)
PRICE (LINEAR LESS)                                              ②
/SUBJECT=SUBJ                                                    ③
/RANK=RANK1 TO RANK15                                           ④
/UTILITY='UTIL.SAV' .                                           ⑤
```

① PLAN 子命令指定 SPSS 外部数据文件 CARPLAN.SAV 作为设计文件，它包括全概念侧面。因为没有 DATA 子命令，工作数据文件就被假定包括这些侧面被访者评价的数据。

② FACTORS 子命令指定因素被期望与秩联系的方法。例如，重量被期望与秩线性相关，则有较大重量的酸奶将得到较低的秩(更被偏爱、更首选)。WARRANTY 因素表明保质期，程序假设它为离散型；MORE 因素表明保质期越长，数值越大，偏爱程度越高；而价格变量即 PRICE 因素与秩之间是线性关系；LESS 表明因素值越高，偏爱程度越低。

③ SUBJECT 子命令指定数据集中的 SUBJ 变量作为标识变量。所有这个变量值相同的观测被认为是同一个观测，程序将它们组合在一起进行效应估计。

④ RANK 子命令指定每个数据点是特定侧面的秩，共有 15 个变量对应这些秩值，而且在包括这些秩的工作数据集中识别这些变量。

⑤ UTILITY 子命令把输出结果写入一个外部数据文件，名为 UTIL.sav，它包括每个被访者的效应估计和有关的统计量。

以上程序表明，一个基本的结合分析程序的结构，应该明确设计文件、调查数据文件的位置或名称。在调查数据文件中一定要有观测的标识变量，否则就会把所有观测当成一个被访者的数据进行处理。

使用 FACTORS 因素指定因素变量虽然可以省略，省略的结果是程序会把设计文件中除 STATUS\_、CARD\_ 变量外都作为因素变量，但是在探讨秩或分数与因素之间的关系时还是不可没有 FACTORS 子命令的。

**【例 7】** 本例仍然使用酸奶偏爱研究的思路。主要注意与 CONJOINT 命令有关的数据来源与文件指定。

```
DATA LIST FREE /CARD_  WARRANTY  WEIGHT  CASING  STATUS_.  ①
BEGIN DATA  ②
1 5 1 600 2
2 5 2 600 2
3 3 2 800 2
4 3 1 1000 2
END DATA.  ③
ADD FILES FILE='/DATA/YUGPLAN.SAV'/FILE=*.  ④
CONJOINT PLAN=*  ⑤
/ DATA='/DATA/YUGDATA.SAV'  ⑥
/ FACTORS= WEIGHT (LINEAR) WARRANTY (MORE)  ⑦
/ SUBJECT=SUBJ /RANK=RANK1 TO RANK15  ⑧
/PRINT=SIMULATION.  ⑨
```

① DATA LIST 定义了 5 个变量：CARD\_ 标识变量、3 个因素变量和 STATUS\_ 变量。  
②、③ BEGIN DATA 和 END DATA 之间的数据是 4 个模拟侧面。每个侧面包括一个 CARD\_ 标识号和感兴趣的因素水平的特殊组合。这 4 个侧面被生成在数据窗中，成为工作数据文件。

所有侧面(观测)的 STATUS\_ 变量值都等于 2。CONJOINT 认为这些 STATUS\_=2 的是模拟侧面。

④ ADD FILES 命令是合并数据文件的过程命令语句。命令后面必须跟“FILE=”指定一个数据文件 YUGPLAN.SAV，用 FIL=\*子命令指定工作数据文件。注意工作数据文件在 ADD FILES 命令的最后指定，所以模拟侧面数据是附加在 YUGPLAN.SAV 的末尾构成新的工作数据集。

⑤ CONJOINT 中的 PLAN 子命令定义这个新的工作数据集作为设计文件。  
⑥ DATA 子命令指定了一个 CONJOINT 要分析的数据文件 YUGDATA.sav。  
⑦ FACTORS 子命令用括号中的参数说明括号前面的变量。指定因素 WEIGHT 重量与秩数据预期是线性关系(LINEAR)；而保质期 WORRANTY 是离散数据(DISCRETE)，但与秩数据的预期关系是保质期越长，数值越大，偏爱程度越高(MORE)。



⑧ SUBJECT 子命令和 RANK 子命令的语句形式和功能与【例 1】相同。

⑨ PRINT 子命令指定只输出模拟观测的分析结果。

**【例 8】** 这是一个有关汽车的偏爱分析研究的例子。因素有 WARRANTY 保质期(1 年、3 年、5 年 3 个水平)、SEATS 座位数(2 座、4 座 2 个水平)、PRICE 价格(7000 美元、10000 美元、14000 美元 3 个水平)、SPED 最高速度(70mile/h、100mile/h、130mile/h)。被访者对设计文件中的 15 个侧面排序, 显然, 秩值为 1~15。秩数据由下面的程序输入、运行, 保存到数据文件中, 设计的 15 个侧面数据也由程序输入、运行, 保存在工作数据文件中。

```
DATA LIST FREE /SUBJ RANK1 TO RANK15. ①
BEGIN DATA ②
01 3 7 6 1 2 4 9 12 15 13 14 5 8 10 11
02 7 3 4 9 6 15 10 13 5 11 1 8 4 2 12
03 12 13 5 1 14 8 11 2 7 6 3 4 15 9 10
04 3 6 7 4 2 1 9 12 15 11 14 5 8 10 13
05 9 3 4 7 6 10 15 13 5 12 1 8 4 2 11
00 12 13 8 1 14 5 11 6 7 2 3 4 15 10 9
END DATA. ③
SAVE OUTFILE='/DATA/RANKINGS.SAV'. ④
DATA LIST FREE /CARD_ WARRANTY SEATS PRICE SPED . ⑤
BEGIN DATA ⑥
1 1 4 14000 130
2 1 4 14000 100
3 3 4 14000 130
4 3 4 14000 100
5 5 2 10000 130
6 1 4 10000 070
7 3 4 10000 070
8 5 2 10000 100
9 1 4 07000 130
10 1 4 07000 100
11 5 2 07000 070
12 5 4 07000 070
13 1 4 07000 070
14 5 2 10000 070
15 5 2 14000 130
END DATA. ⑦
CONJOINT PLAN=* /DATA=' RANKINGS.SAV' ⑧
/FACTORS=PRICE (ANTIIDEAL)SPEED (LINEAR)
WARRANTY (DISCRETE MORE) ⑨
/SUBJECT=SUBJ /RANK=RANK1 TO RANK15. ⑩
```

① 第一个 DATA LIST 定义了 15 个变量, 第 1 个是观测号变量 SUBJ, 其后的 15 个变量是被访者评价的秩, 命名为 RANK1~RANK15。

②、③ BEGIN DATA-END DATA 组产生包含秩的数据文件。

④ 由 BEGIN DATA-END DATA 组的数据产生的文件保存在外部文件 RANKINGS.SAV 中。

⑤ 第二个 DATA LIST 定义的是正交设计中的 4 个因素变量和 1 个 CARD\_变量。

⑥、⑦ 第二个 BEGIN-END DATA 组共 15 个有侧面数据, 没有保存语句跟在后面, 所有运行后的结果在数据窗中, 即作为工作数据文件。

⑧ CONJOINT 命令中, PLAN=\* 使用工作数据文件作为设计文件; DATA 子命令指定了外部数据文件 RANKINGS.sav 作为待分析的数据文件。

⑨ FACTORS 子命令, 假设价格因素 PRICE 水平与秩值是倒二次的关系, 存在一个被认为最差的价格水平, 其他价格水平的偏爱都高于此水平; 假设速度因素 SPEED 水平值与秩值之间是线性关系; 假设保质期是离散的, 保质期越长, 秩值越高, 即更加偏爱。

⑩ SUBJECT 子命令, 定义在数据文件中, 被访者识别号变量为 SUB; RANK 子命令定义秩值变量是 RANK1~RANK15。

上述程序的最后三行⑧~⑩可以写成

```
CONJOINT PLAN=* /DATA=' RANKINGS.SAV'  
/FACTORS=PRICE (ANTIIDEAL) WEIGHT (LINEAR) WARRANTY (DISCRETE MORE)  
/SUBJECT=SUBJ /RANK=RANK1 TO RANK15.
```

需要说明的是:

- ① RANK 子命令指定的数据是按排号顺序安排的侧面的秩。SUBJ 后面的第一个数据点是变量 RANK1, 它是第一个被访者给第一个侧面的秩。
- ② 在设计文件中有 15 个侧面, 所以必须有 15 个秩变量。
- ③ 本例使用了 TO 关键字, 指示有 15 个秩变量。

**【例 9】** 仍然是汽车的偏爱研究, 主要注意 FACTORS 子命令的赋值功能。

```
CONJOINT DATA='DATA.SAV' ①  
/FACTORS=PRICE (LINEAR LESS) WEIGHT (IDEAL 70 100 130)  
WARRANTY (DISCRETE MORE) ②  
/SUBJECT=NO ③  
/RANK=RANK1 TO RANK15. ④
```

① CONJOINT 命令使用 DATA 指定数据文件。它至少应该包括 16 个变量。除了 RANK1~RANK15 外的变量应该是③中 SUBJECT 子命令指定的变量 NO 是观测号。

② FACTOR 子命令指定期望相关。期望价格和秩之间是线性相关, 所以较高的价格偏爱较低(高秩)。期望在速度水平与秩之间是二次相关, 期望较长的保证期与较大的偏爱(低秩)对应。

③ WEIGHT 因素有一个新值列表。如果原来的值为代码 1、2、3, 那么以 70 代替 1, 以 100 代替 2, 以 130 代替 3。

任何设计文件中没有列在 FACTOR 子命令中的变量除了 CARD\_和 STATUS\_外, 都使用 DISCRETE 模型。

## 16.5 结合分析实例

### 16.5.1 课题分析与正交设计

**【例 10】** 本例研究地毯吸尘器的顾客偏爱。

1) 课题内容与因素、因素水平的选择

这是一个流行的结合分析的例题(Green and Wind, 1973)。一个公司对地毯吸尘器的销售感兴趣, 希望调查 5 个对消费者偏爱的影响因素, 包装设计、商标名称、价格、好管家封条、货市的售后保证。经过认真考虑, 包装设计有 3 个水平, 每个水平表明刷子的不同位置; 商标名字有

3 个水平(K2R、Glory、Bissell)；价格有 3 个水平；好管家封条 2 个水平(有、否)；货币的售后保障有 2 个水平(是、否)。表 16-4 所示为在地毯吸尘器研究中的变量名与变量标签。表 16-5 所示是各变量的值和值标签。

表 16-4 变量表

变量名	变量标签
Package	包装设计
brand	商标名称
price	价格
seal	好管家封条
money	货币式售后保证

表 16-5 值和值标签对应表

因素	值	值标签
package	1, 2, 3	A*, B*, C*
brand	1, 2, 3	K2R, Glory, Bissell
price	1.19, 1.39, 1.59	\$1.19, \$1.39, \$1.59
seal	1, 2	是, 否
money	1, 2	是, 否

或许还有其他因素和因素水平描述地毯清洁器，但是对管理者来说只有这些是感兴趣的。对结合分析来说这一点是很重要的。要选择的参与研究的因素必须是认为最影响偏爱的因素变量。使用结合分析，将开发出基于这 5 个因素的顾客偏爱模型。

结合分析的第一步是产生一个展现在被访者面前的产品侧面的因素水平组合。即使很少的因素数和每个因素很少的水平数也会导致一个处理不了的产品侧面的数量。

因此需要产生典型的子集，即正交设计的安排。

生成正交设计程序产生正交安排还涉及正交设计及保存信息到 SPSS 文件中。它不像大多数程序那样，而是在运行该程序之前不必须有一个工作数据集。如果没有工作数据集，则可以选择生成一个，产生变量名、变量标签和在对话框选项中选择值标签。如果已经有了工作数据集，则可以代替它，或者保存正交设计作为一个 SPSS 数据文件。

下面的操作产生正交设计数据。在操作之前数据窗是空的 (SPSS 20.0 不要求)。

2) 正交设计操作过程

(1) 按【数据→正交设计→生成】顺序单击菜单项，见图 16-4，打开【生成正交设计】对话框，见图 16-5。

(2) 定义因素变量名及其标签。

将表 16-6 中的第 1 个变量名 Package 输入【因子名称】框，将其变量标签“包装设计”输入【因子标签】框，单击【添加】按钮，送入按钮旁的矩形框内。在矩形框内显示 Packege'包装设计' (?)。

再将表 16-6 中第 2 个变量名 Brand 和标签“商标名称”分别输入【因子名称】框和【因子标签】框，单击【添加】按钮，送入按钮旁的矩形框内，在矩形框内显示 Brand '商标名称' (?)。依此类推，定义所有变量为因子，并定义它们的标签。具体操作参见第 16.2.3 节的相关内容。

(3) 定义各因素变量的值和值标签。

以定义第一个变量的值和值标签为例说明操作。

在矩形框中选择(单击)第一项 Packege'包装设计'(?), 单击矩形框下面的【定义值】按钮，打开【生成设计：定义值】对话框。在【值】列的 1、2、3 行分别输入 1、2、3，在【标签】列的 1、2、3 行对应位置分别输入值标签 A\*、B\*、C\*。单击【继续】按钮，返回主对话框。

在主对话框中，在第 1 个变量定义的位置显示 Package'包装设计' (1'A\* 2'B\* 3'C\*)。按这样生成一个个定义变量的值标签。

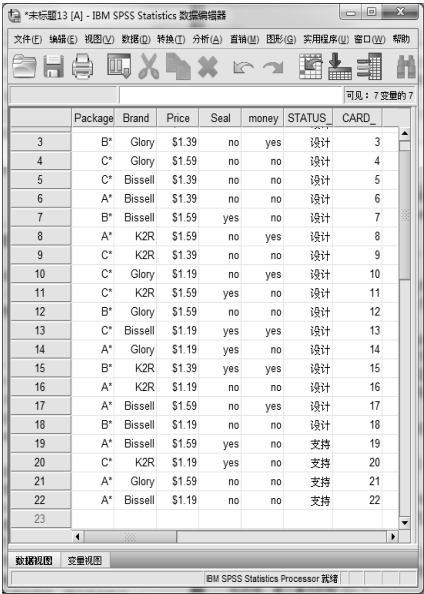
(4) 在主对话框中选择【数据文件】栏中的【创建新数据集】，并在【数据集名称】框中输入“data16-03 吸尘器调查设计”。系统会把设计结果成在一个新的数据窗中。

- (5) 指定随机数种子。在【将随机数初始值重置为】框中输入随机数种子“2000000”。
- (6) 指定生成设计的观测数。单击主对话框中的【选项】按钮，打开如图 16-7 所示的对话框。
- ① 在【生成的最小个案数】框中输入“18”。因为最小观测数是 16，而根据需要，还要求多 2 个侧面，所以输入 18。要求产生 18 种水平组合的侧面 s 数据。
- ② 在【延续个案】栏中选择【延续个案数】，并在其后框中输入保留观测数“4”。
- 单击【继续】按钮，返回主对话框。在主对话框中单击【确定】按钮，提交系统执行。
- (7) 在数据窗中生成设计结果，见图 16-14(a)，其中第 1~18 个观测是设计侧面，第 19~22 是保留侧面；图 16-14(b)比图 16-14(a)中多出 2 个观测，Status\_值为 2，即是模拟侧面。模拟侧面是人为手工输入的数据。

图 16-17 中，变量 STATUS\_值为 0 的是试验侧面，值为 1 的是保留侧面，标签为“支持”。在原设计结果中再输入 2 个模拟观测，其 STATUS\_变量的值是 2，这个数据集名为 A.sav。

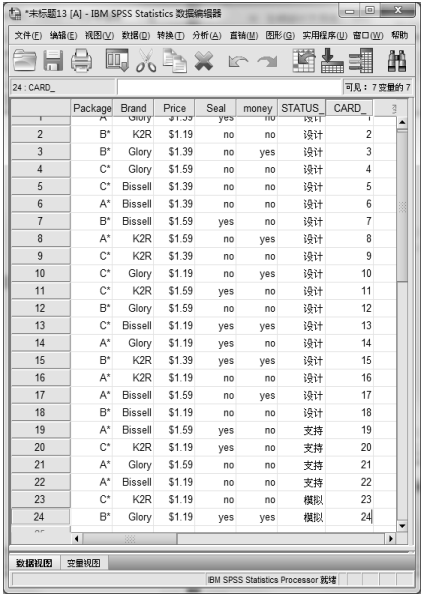
待打印调查卡片和存档列表后，保存为外部文件，名为 data16-03 地毯清洁器调查设计.sav。

(8) 生成设计文件后，还应对所生成的侧面进行查重。如果保留侧面、模拟侧面与试验侧面有重复，再重复上述操作，直到所有侧面没有重复为止。可以说，这种保留侧面与试验侧面重复的现象是偶尔出现的，但是每次试验设计结束后都要查重，保证设计无误。查重使用 DATA 菜单的标识重复个案功能，详见第 2 章有关内容。SPSS 20.0 由系统自动查重，并给出警告信息。



	Package	Brand	Price	Seal	money	STATUS	CARD
3	B*	Glory	\$1.39	no	yes	设计	3
4	C*	Glory	\$1.59	no	no	设计	4
5	C*	Bissell	\$1.39	no	no	设计	5
6	A*	Bissell	\$1.39	no	no	设计	6
7	B*	Bissell	\$1.59	yes	no	设计	7
8	A*	K2R	\$1.59	no	yes	设计	8
9	C*	K2R	\$1.39	no	no	设计	9
10	C*	Glory	\$1.19	no	yes	设计	10
11	C*	K2R	\$1.59	yes	no	设计	11
12	B*	Glory	\$1.59	no	no	设计	12
13	C*	Bissell	\$1.19	yes	yes	设计	13
14	A*	Glory	\$1.19	yes	no	设计	14
15	B*	K2R	\$1.39	yes	yes	设计	15
16	A*	K2R	\$1.19	no	no	设计	16
17	A*	Bissell	\$1.59	no	yes	设计	17
18	B*	Bissell	\$1.19	no	no	设计	18
19	A*	Bissell	\$1.59	yes	no	支持	19
20	C*	K2R	\$1.19	yes	no	支持	20
21	A*	Glory	\$1.59	no	no	支持	21
22	A*	Bissell	\$1.19	no	no	支持	22

(a)



	Package	Brand	Price	Seal	money	STATUS	CARD
1	A*	Glory	\$1.39	yes	no	设计	1
2	B*	K2R	\$1.19	no	no	设计	2
3	B*	Glory	\$1.39	no	yes	设计	3
4	C*	Glory	\$1.59	no	no	设计	4
5	C*	Bissell	\$1.39	no	no	设计	5
6	A*	Bissell	\$1.39	no	no	设计	6
7	B*	Bissell	\$1.59	yes	no	设计	7
8	A*	K2R	\$1.59	no	yes	设计	8
9	C*	K2R	\$1.39	no	no	设计	9
10	C*	Glory	\$1.19	no	yes	设计	10
11	C*	K2R	\$1.59	yes	no	设计	11
12	B*	Glory	\$1.59	no	no	设计	12
13	C*	Bissell	\$1.19	yes	yes	设计	13
14	A*	Glory	\$1.19	yes	no	设计	14
15	B*	K2R	\$1.39	yes	yes	设计	15
16	A*	K2R	\$1.19	no	no	设计	16
17	A*	Bissell	\$1.59	no	yes	设计	17
18	B*	Bissell	\$1.19	no	no	设计	18
19	A*	Bissell	\$1.59	yes	no	支持	19
20	C*	K2R	\$1.19	yes	no	支持	20
21	A*	Glory	\$1.59	no	no	支持	21
22	A*	Bissell	\$1.19	no	no	支持	22
23	C*	K2R	\$1.19	no	no	模拟	23
24	B*	Glory	\$1.19	yes	yes	模拟	24

(b)

图 16-17 地毯清洁器调查设计结果

本例运行后没有出现有重复侧面的警告信息。人为加入的模拟侧面数据已经经过查重，没有重复的观测。

16.5.2 调查准备与调查

- (1) 将设计文件打印成调查用的卡片和存档用的文件。
- ① 按【数据→正交设计→显示】顺序单击菜单项，打开正交设计结果【显示功能】主对话框，见图 16-13。

② 在左侧的变量表中选择 Package、Brand、Price、Seal、Money 这 5 个因素变量，将其移到右侧的【因子】栏中。

③ 在【格式】栏选择两种打印方式：以列表方式或卡片方式输出。

④ 在主对话框中单击【标题】按钮，打开如图 16-14 所示对话框。在【配置文件标题】框内作如下操作：

第 1 行空出，因为回车后第 1 行会有系统默认的标题出现。从第 2 行开始输入自己打印的标题。输入的标题是《地毯清洁剂调查》等。

在【配置文件页脚】框输入提示“请检查所填写的序号是否有重复！”；感谢语“谢谢参与！”

单击【继续】按钮返回主对话框。在主对话框中单击“确定”按钮提交运行。输出结果见图 16-18、图 16-19。

概要文件编号 1: 《地毯清洁剂调查》						概要文件编号 22: 《地毯清洁剂调查》					
请在卡片表格旁填写您对该卡片描述的地毯清洁剂洗好程度排序的结果。						请在卡片表格旁填写您对该卡片描述的地毯清洁剂洗好程度排序的结果。					
卡标识	包装设计	商品名称	价格	好管家封条	货币式售后保 证	卡标识	包装设计	商品名称	价格	好管家封条	货币式售后保 证
1	A*	Glory	\$1.39	yes	no	22	A*	Bissell	\$1.19	no	no
请检查所填写的序号是否有重复。如有重复，请改正！ 谢谢参与！						请检查所填写的序号是否有重复。如有重复，请改正！ 谢谢参与！					

图 16-18 试验侧面与保留侧面卡片例

《地毯清洁剂调查》 请在卡片表格旁填写您对该卡片描述的地毯清洁剂洗好程度排序的结果。						
	卡标识	包装设计	商品名称	价格	好管家封条	货币式售后保 证
1	1	A*	Glory	\$1.39	yes	no
2	2	B*	K2R	\$1.19	no	no
3	3	B*	Glory	\$1.39	no	yes
4	4	C*	Glory	\$1.59	no	no
5	5	C*	Bissell	\$1.39	no	no
6	6	A*	Bissell	\$1.39	no	no
7	7	B*	Bissell	\$1.59	yes	no
8	8	A*	K2R	\$1.59	no	yes
9	9	C*	K2R	\$1.39	no	no
10	10	C*	Glory	\$1.19	no	yes
11	11	C*	K2R	\$1.59	yes	no
12	12	B*	Glory	\$1.59	no	no
13	13	C*	Bissell	\$1.19	yes	yes
14	14	A*	Glory	\$1.19	yes	no
15	15	B*	K2R	\$1.39	yes	yes
16	16	A*	K2R	\$1.19	no	no
17	17	A*	Bissell	\$1.59	no	yes
18	18	B*	Bissell	\$1.19	no	no
19 <sup>a</sup>	19	A*	Bissell	\$1.59	yes	no
20 <sup>a</sup>	20	C*	K2R	\$1.19	yes	no
21 <sup>a</sup>	21	A*	Glory	\$1.59	no	no
22 <sup>a</sup>	22	A*	Bissell	\$1.19	no	no
23 <sup>b</sup>	23	C*	K2R	\$1.19	no	no
24 <sup>b</sup>	24	B*	Glory	\$1.19	yes	yes
请检查所填写的序号是否有重复。如有重复，请改正！ 谢谢参与！						
a. 保留						
b. 模拟						

图 16-19 地毯清洁剂调查试验设计结果列表

(2) 调查。将如图 16-19 所示的整个设计文件列表存档。这是该项目的研究设计结果，前 18 个观测是正交设计结果，还有 4 个保留侧面、2 个模拟侧面。

将图 16-18 所示的卡片打印多份(每份 22 张卡片),用于市场调查。让每个被访者将 22 张卡片认真浏览后,按最喜欢到最不喜欢的顺序排序,将最喜欢的标 1,次之的标 2,依此类推,最不喜欢的标 22。得到数据后,输入数据编辑窗,形成数据文件,见图 16-20。

对每个被访者建立 1 个观测,有 1 个标识号即顺序号;另外 22 个变量分别是被访者对 22 个侧面的排序结果,由于无重复,可以认为每张卡片上标的都是被访者给出的卡片所示侧面的秩。输入数据后的数据窗见图 16-20,保存在数据文件 data16-04 中。

图 16-20 调查数据文件

16.5.3 结合分析编程与结果分析

1. 安排数据文件和设计文件

在数据编辑窗中打开“data16-04 地毯清洁剂调查数据”作为结合分析的数据文件,见图 16-20。设计文件保存在 D:\000SPSS 第 5 版\data 5 版\,名为“data16-03 地毯清洁剂调查设计.sav”。

2. 程序清单

```
CONJOINT PLAN='D:\000SPSS 第 5 版\data 5 版\data16-03 地毯清洁剂调查设计.sav'  
  /DATA=*      /SEQUENCE=PREF1 TO PREF22    /SUBJECT=ID  
  /FACTORS=PACKAGE BRAND (DISCRETE) PRICE (LINEAR LESS)  
           SEAL (LINEAR MORE) MONEY (LINEAR MORE)  
  /PRINT=SUMMARYONLY.
```

3. 程序解释

(1) PLAN 子命令指定设计文件的位置和文件名。

注意:读者在运行程序时需要在 PLAN 子命令中给出自己的文件存储路径,程序方能正常执行。

(2) DATA 子命令指定调查数据文件在数据窗中。

(3) SEQUENCE 子命令指定变量 PREF1~PREF22 分别表示最喜欢的侧面号至最不喜欢的侧面号。例如,第一个观测的 PREF1 的值就是第一个被访者最喜欢的侧面号。

(4) SUBJECT 子命令指定观测的标识变量为 ID。

(5) FACTORS 子命令指定 5 个因素变量是设计文件中除 CARD\_、STATUS\_外的所有变量。指定这些变量与秩的预期关系:价格变量 PRICE、好管家封条变量 SEAL、货币售后保证

MONEY 与秩预期均为线性关系；LESS 表明预期的价格越低秩越低偏爱程度越高；封条是代码，1 代表 no，2 代表 Yes，SEAL 与秩的线性预期参数是 MORE，表明预期被访者更偏爱有好管家封条的产品。同样理解货币售后保证，预期被访者更偏爱有货币售后保证的产品。

(6) PRINT 子命令指定只打印综合分析表。

4. 运行结果(见表 16-6～表 16-15)

5. 结果解释

表 16-6 所示是模型描述，列出了每个变量的水平数和与秩或得分的相关性。表的下面还给出了对试验设计正交性的检验结果。本例的设计经检验为正交设计。

表 16-7 所示是整体效应估计。

表 16-6 模型描述

	水平数	与排列或得分 相关
Package	3	离散
Brand	3	离散
Price	3	线性（小于）
Seal	2	线性（大于）
money	2	线性（大于）

所有因子都是正交因子。

表 16-7 整体效应估计

	实用程序估计	标准误
Package A*	-2.233	.192
B*	1.867	.192
C*	.367	.192
Brand K2R	.367	.192
Glory	-.350	.192
Bissell	-.017	.192
Price \$1.19	-6.595	.988
\$1.39	-7.703	1.154
\$1.59	-8.811	1.320
Seal no	2.000	.287
yes	4.000	.575
money no	1.250	.287
yes	2.500	.575
(常数)	12.870	1.282

如所预期的一样，在价格 Price 与效应之间存在负的关系。较高的价格与较低的效应相关（大的负值意味着较低的效应）。也正如所预期的，封条或货币售后保证与较高的效应相应，有的比没有的效应更高。

表 16-7 中列出了所有因素的各水平的效应，因此可以各因素选择 1 个水平，代表组合成感兴趣的侧面，而这个侧面不一定是在设计中出现的侧面。在表中查出它们的效应，加在一起即给出任何一个组合的总效应。

例如，包装设计为 B\*、商标为 Bissell、价格为 1.59、没有好管家封条和货币式售后保证的清洁器的总效应是

$$\text{utility}(\text{package B*}) + \text{utility}(\text{Bissell}) + \text{utility}(\$1.59) + \text{utility}(\text{no seal}) + \text{utility}(\text{no money-back}) + \text{constant} = 1.867 + (-0.017) + (-8.811) + 2.000 + 1.250 + 12.870 = 9.159$$

如果清洁器的包装设计为 C\*、商标为 K2R、价格 1.39、有认可的好管家封条和货币式售后保证，总效应就是  $0.367 + 0.367 + (-7.703) + 4.000 + 2.500 + 12.870 = 12.401$ 。

这两种地毯清洁器比较，顾客更偏爱后者。

进一步还可以计算顾客最不喜欢的组合和最偏爱的组合，显然最喜欢的组合是包装为 B\*、商标为 K2R、价格为 1.19、有好管家封条和货币式售后保证的。当然，对商家来说，还要与利润一同综合考虑。可能选择的商品是在保证利润的前提下，总效应又比较高的。

表 16-8 所示是相对重要性值。这个表提供了每个因素相对重要性的测度，即重要性分数或重要性值。该值的计算方法是先对每个被访者计算每个因素的效应范围，除以所有因素的效应范围的总和，用百分比表示；再对所有被试者的该因素效应取平均值。重要性值高的因素在顾客看来相对更重要。

如果没有 SUBJECT 子命令，则不对每个被访者进行计算，而是将整个数据文件看作一个被访者计算总效应。重要性计算就像对一个被访者所进行的计算一样。

然而，当使用了 SUBJECT 子命令时，对每个单独的被访者是被平均的，这些平均的重要性将不会与那些使用总效应的计算相一致。

这个结果表明包装设计对整个偏爱最有影响力。这就意味在产品侧面之间存在大的偏爱差异，包括最小的包装要求。结果还表明货币式售后保证在整个决定偏爱中重要性最小。价格是个有重要意义的角色，但是不如包装设计那样大。或许这是因为价格水平之间的差距不是很大。

表 16-9 所示为回归系数。这个表表明 LINEAR 所指定的因素的线性回归系数(程序中没有指定二次模型 IDEAL 和 ANTIIDEAL 模型；如果指定了，或许存在二次项)。特定因素水平的效应由水平与系数相乘来确定。例如，对价格 Price 为\$1.19 的预期效应如表 16-7 所示的-6.595。这是简单地把价格的水平值 1.19，乘以价格系数-5.542 的结果。

表 16-10 所示为相关性检验。表中提供了两个统计量：皮尔逊 R 和肯道尔  $\tau$ ，是观测的和估计参数之间的相关测度。表中还对保留侧面显示了 Kendall's tau。本例有 4 个保留侧面，是由课题决定的没有被结合分析过程使用来估计效应，而结合分析过程对这些侧面计算观测的和预测的秩之间的相关是作为对效应有效性的检验。

表 16-8 重要性值

Package	35.635
Brand	14.911
Price	29.410
Seal	11.172
money	8.872

平均重要性得分

表 16-9 回归系数

	B 系数
	估计
Price	-5.542
Seal	2.000
money	1.250

在许多结合分析中，参数的数量与设计的侧面数关系密切。这会使观测的和估计的分数之间的相关人为地膨胀(增高)。在这种情况下，保留侧面的相关可能给出比较好的对模型拟合的指示。然而要注意，保留侧面将会产生比较低的相关。

表 16-11 和表 16-12 所示是对模拟侧面的分析。模拟侧面是人为输入的两个感兴趣的侧面，不是设计自动生成的，所以在计算估计效应时没有使用这两个侧面。这两个侧面是：

① 包装 package 水平为 C\*，商标 Brand 水平为 K2R，价格 Price 水平为\$1.19，封条和货币式售后保证均为 no。

表 16-10 相关性检验

	值	Sig.
Pearson 的 R	.982	.000
Kendall 的 tau	.892	.000
保留的 Kendall 的 tau	.667	.087

a. 已观测偏好和估计偏好之间的相关性

表 16-11 模拟侧面偏好分数

卡编号	ID	得分
1	23	10.258
2	24	14.292



② 包装 package 水平为 B\*, 商标 Brand 水平为 Glory, 价格 Price 水平为\$1.19, 封条和货币式售后保证均为 Yes。

查表 16-7 中的各因素水平的效应估计值, 相加得到  
模拟观测量①的效应值:  $0.367+0.367-6.595+2+1.25+12.870=10.295$ ;  
模拟观测量②的效应值:  $1.867-0.35-6.595+4+2.5+12.870=14.292$ 。

表 16-12 给出了 3 种模型预测每个模拟侧面可能成为最偏爱的一种属性组合的可能性。  
可以看出, 任何一种模型预测的结果都是第 2 个模拟侧面的概率大于第 1 个模拟侧面。因此选择第 2 个模拟侧面所表达的属性组合作为最偏好的属性组合的可能性最大。

表 16-13 所示为逆相关小结。在 FACTORS 子命令中给出了对 3 个因素的预测模型类型: 对价格的预测是线性模型, LESS 关键字给出预测方向, 即预测被访者对高价格有较低的偏爱; 对封条 Seal 和售后 Money 两个因素的预测是线性模型, 方向是关键字 MORE, 预测对 Yes (②) 比 no (①) 有更高的偏爱。有些被访者给出的秩表明偏爱与所预期的相反, 程序就会对这种情况进行记录和统计。该表显示, 被访者选择与预期相反的情况发生一次的有 3 个被访者, 发生 2 次的有 2 个被访者。

表 16-12 模拟侧面的偏好概率

卡编号	ID	最大效用 <sup>a</sup>	Bradley-Terry-Luce	分对数
1	23	30.0%	43.1%	30.9%
2	24	70.0%	56.9%	69.1%

a. 包括约束模拟  
b. 由于这些主体的得分都是非负数, 因此 Bradley-Terry-Luce 和分对数方法中使用了 10 个主体中的 10 个主体。

表 16-13 逆转摘要

逆转次数	主体数
1	3
2	2

此表显示具有给定逆转次数的主体数。

表 16-14 所示为逆相关统计详表。表明 3 个被访者的选择对价格 Price 因素不是认为越低越好; 2 个被访者的选择, 对货币式售后 Money 因素不是认为有比没有更好; 2 个被访者对封条 Seal 因素不是认为有比没有更好。由于包装 Package、商标 Brand 在 FACTORS 子命令中指定为离散因素, 所以没有逆相关的问题。统计数自然为 0。

表 16-14 逆相关统计

Number of Reversals			
Factor	price		3
	money		2
	seal		2
	brand		0
	package		0
Subject	1	Subject 1	1
	2	Subject 2	2
	3	Subject 3	0
	4	Subject 4	0
	5	Subject 5	0
	6	Subject 6	1
	7	Subject 7	0
	8	Subject 8	0
	9	Subject 9	1
	10	Subject 10	2

在表中的 Subject 部分, 列出了逆相关发生在哪几个被访者身上, 发生了几次。对这几个被访者的回答还可以进行详细研究。

6. 如果程序最后一个语句改变为/PRINT=ALL

数据文件安排与其他语句全部与 5.中所叙述一样。那么输出还会包括对每个被访者数据的一一分析。例如, 对第 5 个被访者的输出, 如表 16-15 所示。

根据计算出的各因素的各水平的效应与标准误, 表 16-15(a)所示是根据第 5 个被访者数据计算出的各因素效应估计值; 表 16-15(b)所示是根据第 5 个被访者数据计算出的各因素重要性值; 表 16-15(c)所示是根据第 5 个被访者数据计算出的 3 个线性模型因素的回归系数; 表 16-15(d)所示是根据第 5 个被访者数据计算出的观测量的和估计参数之间的相关测度。两个统计量为皮尔逊 R 和肯道尔  $\tau$ ; 表 16-15(e)所示是根据第 5 个被访者数据计算出的 2 个模拟侧面的偏爱分数。

表 16-15 对每个被访者数据的分析输出(以 ID=5 为例)

		实用程序估计	标准误
Package	A*	-6.000	.313
	B*	3.000	.313
	C*	3.000	.313
Brand	K2R	.167	.313
	Glory	-1.000	.313
	Bissell	.833	.313
Price	\$1.19	-19.833	1.614
	\$1.39	-23.167	1.886
	\$1.59	-26.500	2.157
Seal	no	1.000	.470
	yes	2.000	.940
money	no	1.000	.470
	yes	2.000	.940
(常数)		30.000	2.095

(a)

Package	46.154
Brand	9.402
Price	34.188
Seal	5.128
money	5.128

(b)

	B 系数	
	估计	标准误
Price	-16.667	1.357
Seal	1.000	.470
money	1.000	.470

(c)

	值	Sig.
Pearson 的 R	.991	.000
Kendall 的 tau	.957	.000
保留的 Kendall 的 tau	1.000	.021

a. 已观测偏好和估计偏好之间的相关性

(d)

卡编号	ID	得分
1	23	15.333
2	24	16.167

(e)

习 题 16

1. 对于市场调查中顾客偏爱的分析必须要用结合分析吗？
2. 结合分析适用于什么样的数据？主要解决什么问题？
3. 要调查分析某产品不同侧面组合的顾客偏爱，整个工作要分哪几个主要步骤？每个步骤可以用 SPSS 的哪些程序解决？每个步骤的作用是什么？
4. 使用 Conjoint 命令语句编程，必须包括什么语句？
5. 如果数据窗中有试验设计数据，那么程序中可以减少哪个语句？
6. 市场上先调查了解市民曾购买的酸奶有几种，主要因素有重量等级、品牌、价格、保质期。每个因素取 2~3 个等级，设计使用结合分析了解市民偏爱的课题解决方案。如果可能，将调查数据使用程序分析并得出结论。
7. 设计一个台式个人计算机的顾客偏爱课题及其解决方案。

# 第 17 章 时间序列分析

时间序列是指依时间顺序取得的观察资料的集合。在一个时间序列中，离散样本序列可以按相等时间间隔或不相等时间间隔获取，更多的是采用前者来实现。时间序列的特点是数据资料的先后顺序不能随意改变，逐次的观测值通常是不独立的，而且分析时必须考虑观测资料的时间顺序，这同以前所介绍的观测资料有很大的区别。

时间序列的变化受多种因素的影响，一般可将这些因素分为以下 4 种：

(1) 长期趋势( $T$ )。

长期趋势反映了某种现象在一个较长时间内的发展方向，可以在一个相当长的时间内表现出一种近似直线的持续向上、持续向下或平稳的趋势，也可表现出某种类似指数趋势或其他曲线趋势。粗略地可将“趋势”定义为“均值的长期变化”。Granger(1966)定义“均值趋势”为包含波长超过观测时间序列长度的所有频率分量。长期趋势一旦形成，便会延续很长时间，因此对其进行预测研究具有特别重要的现实意义。

(2) 季节变动( $S$ )。

季节变动是某种现象受季节变动影响所形成的一种长度和幅度固定的周期波动。许多时间序列如销售量及温度等都显示出年周期的变化。

(3) 周期变动( $C$ )。

周期变动也称循环变动，它是由于某些物理原因或经济原因的影响而显示出有固定周期的变化。例如，股票价格的变化等具有明显的周期变动特征。

周期变动有时具有季节变动的特征，如像季节变动一样可以预计它缓慢地上下波动，但这里的“周期”一词是用来描述比季节变动更难以预测、更加缓慢的移动。周期长度和峰值都是不确定的，许多周期的平均长度为 3~4 年，有的达 15 年以上。一些学者长期以来一直致力于研究周期的本质和可测性。

对于短期预测，通常会将周期和趋势放在一起考虑，因为此时不可能从短期序列中获取任何有关周期的有用信息。

(4) 不规则变动( $I$ )。

不规则变动因素又称随机变动，它是受各种偶然因素的影响所形成的不规则波动，如石油价格受突发事件的影响上涨等。

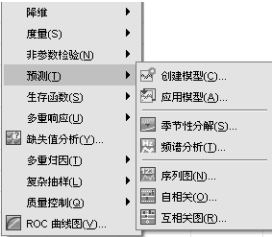
当将时间序列分解成长期趋势、季节变动、周期变动和不规则变动 4 个因素后，可以将时间序列  $Y$  看成这 4 个因素的函数，即  $Y_t = f(T_t, S_t, C_t, I_t)$ 。

常用的时间序列分解的模型有加法模型和乘法模型。加法模型为  $Y_t = T_t + S_t + C_t + I_t$ ；乘法模型为  $Y_t = T_t \times S_t \times C_t \times I_t$ 。

相对而言，乘法模型比加法模型用得更多。在乘法模型中，时间序列值和长期趋势用绝对值表示，季节变动、周期变动和不规则变动用相对值(百分数)表示。

本章主要介绍时间序列分析研究中的序列图、建立模型(指数平滑、综合移动平均)、应用模型、自相关、季节分解、频谱分析、互相关等时间序列分析方法及程序的使用。

SPSS 中进行时间序列分析由主菜单的【分析】下拉菜单中的【预测】菜单项导出，见图 17-1，其中包括：



- 创建模型；
- 应用模型；
- 季节性分解；
- 频谱分析；
- 序列图；
- 自相关；
- 互相关图。

图 17-1 各种时间序列分析过程

## 17.1 时间序列的建立和平稳化

由于大多数时间序列模型需要完整的数据序列，因此时间序列中不能含有缺失值；只有按 SPSS 中的要求为时间序列建立时间变量，分析时对应的序列才能被 SPSS 识别为时间序列；而只有基于平稳化的时间序列，才能有效地开展进一步的分析。

因此在选择上述过程对数据用时间序列模型进行拟合处理前，应先对数据进行必要的预处理。预处理分为三个步骤，首先对有缺失值的数据进行修补，其次将数据资料定义为相应的时间序列，最后对时间序列数据平稳性。

如果数据文件中存在一个变量，其值是按某一时间间隔采集的，要进行时间序列分析，还需要有一个表明采集时间的日期变量。生成日期变量的方法详见 2.1.3 节中的相关内容。

### 17.1.1 缺失值数据的替换

如果要进行时间序列分析的数据存在缺失值，就不能采用通常删除的办法来解决，因为这样会导致原有时间序列周期性的破坏，而无法得到正确的分析结果。

替换缺失值可在【转换】菜单的【替换缺失值】过程中进行。按【转换→替换缺失值】顺序单击菜单项，打开【替换缺失值】对话框，见图 17-2。

(1) 从源变量表中选择需替换缺失值的变量，将其送入【新变量】框。

(2) 在【名称】框中存储替换缺失值后时间序列的新变量名。

(3) 【方法】下拉列表中提供了替换缺失值的方法。共有 5 种，分别是：

①【序列均值】。用整个序列的均数来替换缺失值，这是系统默认选项。

②【临近点均值】。用相邻若干点的有效值的均数替换缺失值，在【附(临)近点的跨度】框中输入计算均数使用的相邻点数。

③【临近点中位数】。用若干相邻点的中位数替换缺失值，在【附(临)近点的跨度】框中输入计算中位数使用的相邻点数。



图 17-2 【替换缺失值】对话框

④【线性插值法】。用相邻两点的平均值替换缺失值。如果时间序列的最前或最后数据有缺失值，则缺失值不被替换。

⑤【点处的线性趋势】。用该点的线性趋势替换缺失值。将记录号作为自变量，时间序列值作为因变量进行回归，求得该点的预测值替换缺失值。

(4)【附(临)近点的跨度】框。设置上述相应替换方法中需要使用的相邻点数。输入大于等于 2 的整数。如果用时间序列中全部的有效值，选择【全部】。

(5)【更改】按钮。若替换方法有变化，单击该按钮可将所作的修改应用于相应的变量。

设置完成后，单击【确定】按钮运行。

## 17.1.2 建立时间序列新变量

时间序列分析是建立在序列平稳的条件上的，判断序列是否平稳可以看它的均数和方差是否不再随时间的变化而变化，自相关系数值是否只与时间间隔有关而与所处的时间无关。大多数时间序列是不平稳的。因此，首先要识别并将不平稳的时间序列变成平稳的时间序列。

为获取平稳的时间序列，经常要使用一阶差分、二阶差分。有时为选择一个合适的时间序列的模型，还要对原时间序列数据进行对数转换或平方根转换等。这就需要在已经建立了时间序列的数据文件中，再建一个新的时间序列的变量。在 SPSS 中创建时间系列可根据现有的数值型时间序列变量的函数建立一个新变量。所建的这些转换值在许多时间序列的分析程序中经常用到。

判定时间序列的平稳性和趋势特征，可借助于各种图形，如序列图、自相关图、频谱图等。

### 1. 建立时间序列新变量的方法

(1) 在 SPSS 的主菜单中，按【转换→创建时间序列】顺序单击菜单项打开对话框，见图 17-3。

(2) 选择一个用来建立新变量的数值型变量，单击向右箭头按钮，在【变量→新名称】框中出现等式，等号左边是默认的新变量名，右边是一个在【函数】下拉列表中选定的转换函数，函数中的参数就是选定的需转换的变量名。

(3) 在【名称和函数】栏的【名称】框中，出现默认的新变量名，由建立它的变量名的前 6 个字符接下画线和 1 个有序数字组成。例如，变量 sales 对应的新变量名是 sales\_1。如果要自定义变量名，可输入自定的变量名，单击【更改】按钮确认，【更改】按钮有两个作用。其一是自定义新变量名后单击该按钮，可以改变新变量名为自定义名；其二是选择函数以后单击该按钮，在上面的【变量→新变量】框中的新变量名前的转换函数就不是系统默认的函数了，而是选择的函数。新变量保持原变量的值标签。

(4)【函数】下拉列表中提供的有效函数如下：

①【差值(分)】函数。计算时间序列里相邻值之间的非季节性差分。【顺序】(阶数)框中的数用来计算差分的样品之前的样品数。因为计算一次差分就会丢失一个观测。如果【顺序】框中的数为  $n$ ，则新时间序列变量中开始的  $n$  个值是系统缺失值。例如，【顺序】框中输入“2”，则新变量前 2 个样品将成为系统缺失值。



图 17-3 【创建时间序列】对话框

②【季节性差分】函数。恒定跨距的序列值之间的差值。跨距取决于当前定义的周期。要计算季节差,必须已经定义了(【数据→定义日期】)包括周期成分的日期变量(如该年的月份)。  
【顺序】框中的数是用于计算差值的季节周期,在时间序列开始的系统缺失值的观测数等于周期乘以【顺序】框中的数值。例如,如果当前周期是 12,【顺序】框中的数值是 2,那么新变量前 24 个值将是系统缺失值。

③【中心移动平均】函数。以当前值为中心,在指定跨距范围内计算包括当前值的序列值的平均值。跨距是用于计算平均值的个数。如果跨距是偶数,则移动平均数就是每对非中心值的均值再与当前值一起求得的均值。例如,如果跨距为 1,则中心移动平均值就是当前值;如果跨距为 3;则中心移动平均值就是当前值与其前后各 1 个值的均值;如果跨距为 4,则中心移动平均值就是当前值前后两对值(非中心点的 4 个值)的均值再与当前值一起求得的均值。如果跨距为  $n$ ,在新变量序列开始和结尾处的系统缺失值数就等于  $n/2$  的整数。例如,跨距是 5,则计算中心移动平均值的中心点为第 3 个观测处,因此,在开始和结尾处有系统缺失值的观测数是 2。

④【先前移动平均】函数。计算当前值之前的指定跨距中的观测值的平均值。跨距是用来计算平均值的前面时间序列值的数量,该序列开始处缺失值的数量等于跨距。

⑤【运行(移动)中位数】函数。计算以当前值为中心,在指定跨距范围内(包括当前值)的序列值的中位数。计算中位数的时间序列值的数量称为跨距。如果跨距是偶数,中位数是每对非中心观测值的中位数的平均值。对偶数跨距的值和奇数跨距的值而言, $n$  跨距时间序列起始位置和结束位置含有系统缺失值的样品的数目等于  $n/2$  的整数。例如,如果跨距是 5,则系统缺失值在时间系列的开始和结束处各有 2 个。

⑥【累计求和】函数。新序列值为原时间序列截止到当前值时的累计和。

⑦【滞后】函数。产生滞后序列,即将前  $k$  时点的值作为当前值。 $K$  为指定滞后的阶数,它是当前样品之前的样品的数量。在新的时间序列的开始处含有系统缺失值的样品数等于【顺序】框中设定的值。

⑧【提前(领先)】函数。产生提前序列,即将后  $k$  时点的值作为当前值。 $K$  为指定领先的阶数,它是当前样品之后的样品的数量。在新的时间序列的末端含有系统缺失值的样品数等于【顺序】框中设定的值。

⑨【平滑】函数。用它可以计算原序列的 T4235 平滑序列,该法又称为 T4253H 平滑法,最早由 Tukey 提出。对经 T4235 平滑处理后得到的序列用 Hanning 权重求移动平均,从而得到新序列,故新序列是建立在复合数据平滑法基础上的。它的功能是通过多步处理将序列中的异常值剔除,使序列平滑。Velleman 于 1980 年给出如下算法。

设原始序列为  $X_t, t=1,2,\dots,n$ , 首先计算窗宽为 4 的中位数。

令  $Z$  为平滑序列, $Z$  的下标表示中位数所在时点的位置,上标表示不同计算的阶段,则

$$Z_{(j+1)/2} = \text{median}(X_{j-1}, X_j, X_{j+1}, X_{j+2}) \quad (j=1,2,\dots,n-2)$$

并且

$$Z_{0.5} = X_1$$

$$Z_{1.5} = \text{median}(X_1, X_2) = (X_1 + X_2) / 2$$

$$Z_{(n-1)/2} = \text{median}(X_{n-1}, X_n) = (X_{n-1} + X_n) / 2$$

$$Z_n = X_n$$

然后,对序列  $Z$  计算窗宽为 2 的中位数

$$Z_1^{(1)} = Z_{0.5}, Z_n^{(1)} = Z_{(n+1)/2}$$

并且 
$$Z_j^{(1)} = \frac{1}{2}[Z_{(j-1)/2} + Z_{(j+1)/2}] \quad (j = 2, 3, \dots, n-1)$$

接着, 对前一步中的  $Z_1^{(1)}, \dots, Z_n^{(1)}$  计算窗宽为 5 的中位数, 得到  $Z^{(2)}$

$$Z_1^{(2)} = Z_1^{(1)}$$

$$Z_n^{(2)} = Z_n^{(1)}$$

$$Z_2^{(1)} = \text{median}(Z_1^{(1)}, Z_2^{(1)}, Z_3^{(1)})$$

$$Z_{n-1}^{(2)} = \text{median}(Z_{n-2}^{(1)}, Z_{n-1}^{(1)}, Z_n^{(1)})$$

并且 
$$Z_j^{(2)} = \text{median}(Z_{j-2}^{(1)}, Z_{j-1}^{(1)}, Z_j^{(1)}, Z_{j+1}^{(1)}, Z_{j+2}^{(1)}) \quad (j = 3, 4, \dots, n-2)$$

继续对前一步中的  $Z_1^{(2)}, \dots, Z_n^{(2)}$  计算窗宽为 3 的中位数, 得到  $Z^{(3)}$

$$Z_j^{(3)} = \text{median}(Z_{j-1}^{(2)}, Z_j^{(2)}, Z_{j+1}^{(2)}) \quad (j = 2, 3, \dots, n-1)$$

$$Z_1^{(3)} = \text{median}(3Z_2^{(3)} - 2Z_3^{(3)}, Z_1^{(3)}, Z_2^{(3)})$$

$$Z_n^{(3)} = \text{median}(3Z_{n-1}^{(3)} - 2Z_{n-2}^{(3)}, Z_n^{(3)}, Z_{n-1}^{(3)})$$

最后, 对序列  $Z_1^{(3)}, \dots, Z_n^{(3)}$  进行 Hanning 加权修匀, 得到

$$Z_j^{(4)} = \frac{1}{4}Z_{j-1}^{(3)} + \frac{1}{2}Z_j^{(3)} + \frac{1}{4}Z_{j+1}^{(3)} \quad (j = 2, 3, \dots, n-1)$$

$$Z_1^{(4)} = Z_1^{(3)}$$

$$Z_n^{(4)} = Z_n^{(3)}$$

残差为

$$D_i = X_i - Z_i^{(4)} \quad (i = 1, 2, \dots, n)$$

对残差  $D_1, \dots, D_5$  重复前面的步骤, 则得到最终结果  $D_1^{(4)}, \dots, D_n^{(4)}$ 。

因此, 最终平滑的结果为

$$Y_i = Z_i^{(4)} + D_i^{(4)} \quad (i = 1, 2, \dots, n)$$

(5) 单击【确定】按钮, 提交系统运行, 可在输出窗和原数据窗中看到运行结果。原数据窗中产生新变量序列, 输出窗中有转换小结。

## 2. 建立时间序列新变量实例

【例 1】数据文件 data17-01 为某公司 1973—1999 年的销售额(万元)。用【滞后】函数建立新变量。

(1) 按【转换→创建时间系列】顺序打开【创建时间系列】对话框, 见图 17-3。

(2) 在【函数】下拉列表中选择【滞后】函数。选择 sales 变量, 将其移到【变量→新名称】框中。

(3) 【名称和函数】框的【名称】框中是默认的新变量名 sales\_1。单击【确定】按钮。

输出如表 17-1 所示。在工作的数据文件中生成 sales\_1 滞后新变量。

在表 17-1 中, 第一行从左至右各列依

表 17-1 运行滞后函数时的结果说明

创建序列				
	序列名	非缺失值的个案数		有效个案数
		第一个	最后一个	
1	sales_1	2	27	26
LAGS(sales, 1)				

次显示的为：新序列变量名、第一个非缺失值观测号、最后一个非缺失值观测号、有效样品数、创建新序列使用的函数。

## 17.2 序 列 图

序列图是时间序列的基本观察工具。在构建一个模型以前，为了了解数据的性质、数据是否有季节性波动，可以通过对时间序列绘制连续的样品图来加以判断。序列图是线图的一种特殊形式，它以时间变量为横轴，以分析变量为纵轴，并对线图作了一些加工，比一般线图有更多适合时间序列特点的功能。



图 17-4 【序列图】主对话框

### 17.2.1 序列图过程

按【分析→预测→序列图】顺序单击菜单项，打开如图 17-4 所示的【序列图】主对话框。

#### 1. 定义变量

在源变量表中选定一个或多个满足时间序列要求的或是按有意义顺序排序样品的变量，移到【变量】框中。

#### 2. 定义时间轴标签变量

在源变量表中选择一个分类变量，移到【时间轴标签】框中。这个变量可以是数值型、字符型或长字符型变量。该变量的值用来标示时间轴。

#### 3. 数据转换

【转换】栏中提供了 3 种对时间序列或类似数据进行转换的方法。

- (1) 【自然对数转换】。用数据值的自然对数来代替它们本身。这种转换需要所有的值大于 0。
- (2) 【差分】。计算两个相邻变量之间的差值。输入一个正整数作为差分的阶。一阶差分是用当前值减前一个值。二阶差分是对一阶差分序列作同样的处理，而不是采用每个值减其之前两个样品的那个值。
- (3) 【季节性差分】。通过计算两个时间跨度相同的序列值之间的差值来转换时间序列数据。输入一个正整数作为计算差值的时间周期数。这种转换只有当序列的周期已经定义过时才是有效的(使用【数据】菜单中【定义日期】对话框定义)。

#### 4. 选择图形的输出方式

选择【每个变量对应一个图表】，为【变量】框中的每个变量产生一张图。若不选择本选项，则所有的变量绘制在同一张图上。

#### 5. 定义时间轴基准线

单击【时间线】按钮，打开如图 17-5 所示的【时序图：时间轴参考线】对话框，对绘制的直方图、线图或散点图选择一条任意的刻度线或分类轴线进行定义。



(1) 【无参考线】。输出的图形中没有基准线。

(2) 【每一个更改的线】。基准线会随参考变量的改变而变化。左侧的变量列表显示了数据文件中未在主对话框中指定的变量，在其中选择一个变量并移到【参考变量】框中作为参考变量。

(3) 【在日期上的线】。在一个特定的点显示用日期或观察数定义的单条参考线。

① 如果已经定义了日期，显示用【定义日期】定义的所有日期的标识部分。输入想要显示参考线处的日期。

② 如果没有定义日期，则输入想要显示参考线处的参考变量的值。

## 6. 定义时间轴的格式

在主对话框中单击【格式】按钮，打开如图 17-6 所示的【时序图：格式】对话框。选择作图类型以及有关的格式参数。

(1) 【水平轴上的时间】。要求水平轴是时间轴，用垂直轴表示序列值。

(2) 【单个变量图】栏。当只选了一个变量，或选择了【每个变量对应一个图表】时，则对指定的绘图变量可选择线【线图】或【面积图】；选中【序列均值的参考线】则在序列均值处画参考线。



图 17-5 【时序图：时间轴参考线】对话框



图 17-6 【时序图：格式】对话框

(3) 【多个变量图】栏的【连接变量之间的个案】。一张图中显示多个变量的序列图，以示变量间和各观察值之间的联系。

单击【继续】按钮，返回主对话框。单击【确定】按钮，运行绘制序列图程序。

## 17.2.2 序列图应用实例

【例 2】 data17-02.sav 是 SPSS16.0 中自带的假设数据文件，包括 1999 年 1 月至 2003 年 12 月 4 年中 85 个地区宽带供货商每月的国家宽带服务用户数量的数据。试用总用户数量序列作序列图。

该数据文件中，market-1—market-85：变量标签分别为供货商 1 的用户数至供货商 85 的用户数；total：变量标签为总用户数量；它们均为数值型的尺度测度变量。

具体操作步骤如下：

(1) 按【分析→预测→序列图】顺序单击菜单项，打开如图 17-4 所示的主对话框。

(2) 在源变量框中选“总用户数量”变量作为绘图变量，移到【变量】框中。

- (3) 在源变量框中选择“Date 变量”作为时间轴变量，移入【时间轴标签】框中。
- (4) 单击【时间线】按钮，在【时序图：时间轴参考线】对话框中。选择【在日期上的线】，在【年】框中输入 2002，在【月】框中输入 6。单击【继续】按钮返回【序列图】主对话框。
- (5) 其他使用系统默认选项，单击【确定】按钮提交运行，在输出窗中得到如表 17-2、表 17-3 和图 17-7 所示的结果。

表 17-2 模型描述表

模型描述	
模型名称	MOD_4
序列或顺序 1	总用户数量
转换	无
非季节性差分	0
季节性差分	0
季节性期间的长度	12
水平轴标签	Date_
干预开始	YEAR, not periodic=2002, MONTH, period 12=6
参考线	无
曲线下方的区域	未填充

正在应用来自 MOD\_4 的模型指定。

表 17-3 样品处理摘要

个案处理摘要	
序列或顺序长度	总用户数量
图中的缺失值数	63
用户缺失	0
系统缺失	3

(6) 结果解释。表 17-2 中给出了模型的描述，从上至下依次显示的是模型名 MOD\_1、序列或顺序 1(名称为“用户数量”)、转换(无)、非季节性差分(0)、季节性差分(0)、季节性期间的长度(12)、水平轴标签(Date)、插入参考线的起始时间(2002 年 6 月)、参考线(无)、曲线下方的区域(未填充)。括号中是第二列中对应各行的结果。

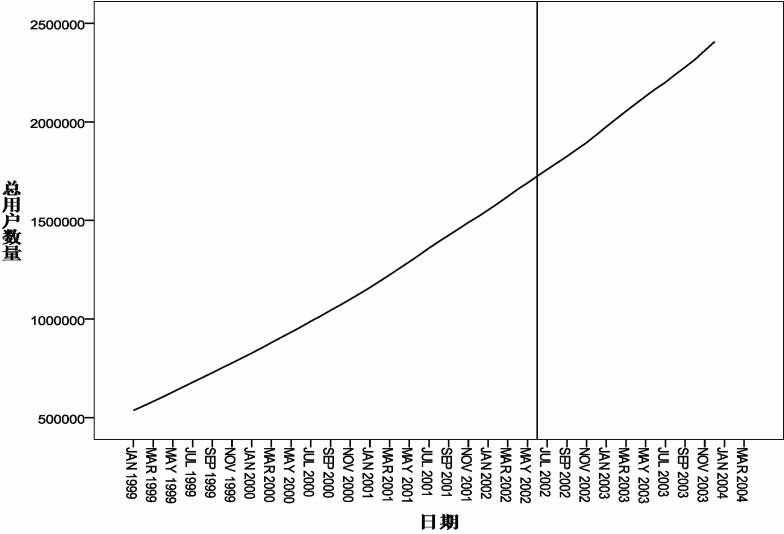


图 17-7 含有参考线的序列图

表 17-3 中给出了样品处理的摘要。从上至下依次显示：序列或顺序长度(63，缺失值数量：用户缺失值(0)、系统缺失值(3))。

图 17-7 所示为供货商 1 的用户总数的序列图。图中竖线为参考线，对应的时间为 2002 年 6 月。由图可见，序列展现很平滑的向上趋势，没有一点季节性波动。总的来说，季节性变化趋势不是数据的显著特征。

当然，如果要对总用户数以外的其他序列作时间序列分析，在排除有季节性模型的可能性前，应分别检查各个序列。

## 17.3 建立时间序列模型

时间序列建模程序(TSMODEL)可为单变量时间序列构建指数平滑、自回归综合移动平均 (ARIMA) 和传递函数(TF) 模型并产生预报。程序包括为每个因变量时间序列自动识别和估计适合模型的专家建模器，因此，不需要通过反复试验来识别一个适当的模型。也可以用自定义的方式来指定一个模型。这个程序的设计得到芝加哥大学 Ruey Tsay 教授的帮助。

按【分析→预测→创建模型】顺序单击菜单项，打开如图 17-8 所示的建模提示框。

**注意：**在使用该对话框之前，应对时间序列定义起始时间和时间间隔，以确保输出标识正确，并且如果需要，可获得季节模型。

下次运行时，若不需要显示该提示框，可选择【不再显示此消息】。

单击【定义日期】按钮，打开【定义日期】对话框，可设置起始时间和时间间隔。具体操作方法参见第 2 章相关内容。

如果当前的数据文件已经定义为时间序列时，该对话框不出现。

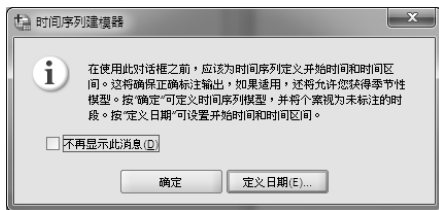


图 17-8 【时间序列建模器】提示框

### 17.3.1 指数平滑与 ARIMA 模型概述

#### 1. 指数平滑

指数平滑预测方法最先由 C. C. Holt 在 1958 年提出，它最初只应用于无趋势、非季节作为基本形式的时间序列的分析，后经 Brown、Winter 等统计学家的深入研究和发 展，指数平滑涉及的数据内部构成更丰富，相应的数据处理方法也更多。指数平滑法的估计是非线性的，其目标是使预测值和实测值间的均方差(MSE)最小。

指数平滑法的基本计算公式如下：

$$\hat{Y}_{t+1} = \frac{\sum_{j=0}^{\infty} \theta^j Y_{t-j}}{\sum_{j=0}^{\infty} \theta^j} = (1 - \theta) \sum_{j=0}^{\infty} \theta^j Y_{t-j}$$

式中， $Y_t$  表示观测序列； $\hat{Y}_t$  表示预测序列，分母为正则化常数，称  $\theta$  为平滑参数，其作用是保证权重之和为 1； $0 \leq \theta \leq 1$ ； $j=0, 1, 2, \dots$ ； $t=1, 2, \dots$ ； $(t > j)$ 。

在 SPSS 中，提供了 7 种方法来处理时间序列中的随机波动、长期趋势及周期性波动的指数平滑模型。其中，无季节性的指数平滑模型有 4 种，季节性的指数平滑模型有 3 种。

在指数平滑法中,  $\alpha$  表示水平平滑权重,  $\gamma$  表示趋势平滑权重,  $\phi$  表示阻尼趋势平滑权重,  $\delta$  表示季节平滑权重。设在调查研究中得到的单变量时间序列为  $Y_t(t=1,2,\cdots,n)$ , 总观测值数量为  $n$ , 序列  $Y$  在时间  $t$  时由模型估计的领先  $k$  步的预测值为  $\hat{Y}_t(K)$ , 季节长度为  $s$ 。

(1) 无季节性的指数平滑模型。

① 简单指数平滑法(Simple)。简单指数平滑法只有单个水平参数, 并可用下面的等式来描述:

$$\begin{aligned} L(t) &= \alpha Y(t) + (1-\alpha)L(t-1) \\ \hat{Y}_t(K) &= L(t) \end{aligned}$$

式中,  $L(t)$  为本期预测值;  $L(t-1)$  为前一期预测值。用第一期的实际值或最初  $k$  期的观测值的平均值作为其初始值。其功能等价于 ARIMA(0,1,1) 过程。

② Holt 指数平滑法。Holt 指数平滑法有水平参数和趋势参数, 并可用下面的等式来描述:

$$\begin{aligned} L(t) &= \alpha Y(t) + (1-\alpha)[L(t-1) + T(t-1)] \\ T(t) &= \gamma[L(t) - L(t-1)] + (1-\gamma)T(t-1) \\ \hat{Y}_t(k) &= L(t) + kT(t) \end{aligned}$$

式中,  $T(t-1)$  为前一期的趋势值;  $T(t)$  为本期趋势值;  $k$  为超前期数, 其余同上。其功能等价于 ARIMA(0,2,2) 过程。

③ Brown 指数平滑法。Brown 指数平滑法有水平参数和趋势参数, 并可用下面的公式来描述:

$$\begin{aligned} L(t) &= \alpha Y(t) + (1-\alpha)L(t-1) \\ T(t) &= \alpha[L(t) - L(t-1)] + (1-\alpha)T(t-1) \\ \hat{Y}_t(k) &= L(t) + [(k-1) + \alpha^{-1}]T(t) \end{aligned}$$

其功能等价于 ARIMA(0,2,2) 过程, 并受到 MA 中参数的约束。

④ 阻尼趋势指数平滑法。阻尼趋势指数平滑法有水平参数和阻尼趋势参数, 并可用下面的公式来描述:

$$\begin{aligned} L(t) &= \alpha Y(t) + (1-\alpha)L(t-1) + \phi T(t-1) \\ T(t) &= \gamma[L(t) - L(t-1)] + (1-\gamma)\phi T(t-1) \\ \hat{Y}_t(k) &= L(t) + \sum_{i=1}^k \phi^i T(t) \end{aligned}$$

其功能等价于 ARIMA(1,1,2) 过程。

(2) 季节性指数平滑模型。

① 简单季节性指数平滑法。简单季节性指数平滑法有水平参数和季节性参数, 并可用下面的公式来描述:

$$\begin{aligned} L(t) &= \alpha[Y(t) - S(t-s)] + (1-\alpha)L(t-1) \\ S(t) &= \delta[Y(t) - L(t)] + (1-\delta)S(t-s) \\ \hat{Y}_t(k) &= L(t) + S(t+k-s) \end{aligned}$$

式中,  $S(t)$  称为季节修正系数。其功能等价于 ARIMA[0,1,(1, $s$ , $s+1$ )](0,1,0) 过程, 受 MA 中参数的约束。

② Winters 加性指数平滑法。Winters 加性指数平滑法有水平、趋势和季节性参数, 并可

用下面的公式来描述:

$$L(t) = \alpha[Y(t) - S(t-s)] + (1-\alpha)[L(t-1) + T(t-1)]$$

$$T(t) = \gamma[Y(t) - L(t-1)] + (1-\gamma)T(t-1)$$

$$S(t) = \delta[Y(t) - L(t)] + (1-\delta)S(t-s)$$

$$\hat{Y}_t(k) = L(t) + kT(s) + S(t+k-s)$$

其功能等价于 ARIMA(0,1,s+1) (0,1,0) 过程,受到 MA 中参数的约束。

③ Winters 积性指数平滑法。Winters 积性指数平滑法有水平、趋势和季节性参数, 并可下面的公式来描述:

$$L(t) = \alpha[Y(t) / S(t-s)] + (1-\alpha)[L(t-1) + T(t-1)]$$

$$T(t) = \gamma[L(t) - L(t-1)] + (1-\gamma)T(t-1)$$

$$S(t) = \delta[Y(t) / L(t)] + (1-\delta)S(t-s)$$

$$\hat{Y}_t(k) = [L(t) + kT(s)]S(t+k-s)$$

没有等价的 ARIMA 模型。

## 2. ARIMA 模型和传递函数模型

ARIMA 自回归综合移动平均模型和 TF 传递函数模型是广泛应用于时间序列分析的常见模型。ARIMA 模型就是著名的 Box-Jenkins 模型。它可以延伸到对包含季节趋势的时间序列进行分析。根据对时间序列特征的预先研究, 可以指定 3 个参数用来分析时间序列, 即自回归阶数( $p$ )、差分次数( $d$ )和移动平均阶数( $q$ )。通常模型被写作 ARIMA( $p,d,q$ )。

Box-Jenkins 方法的第一步是对时间序列数据求一阶差分  $\nabla X_t = X_t - X_{t-1}$ 、二阶差分  $\nabla^2 X_t = \nabla X_t - \nabla X_{t-1} \cdots$  直到它是平稳序列为止。式中,  $\nabla$  为一阶差分算子,  $\nabla^2$  为二阶差分算子。这可以通过检查各种差分序列的相关图(包括偏自相关图)直到找出一个“急速”下降于零, 并且从此任何季节效应都大大消除的序列来完成对时间序列的随机性、平稳性及季节性的分析。对于非季节数据, 通常求一阶差分就足够了。对周期为 12 的季节数据, 如果季节效应是加性的, 通常可以采用算子  $\nabla_{12}$ ; 如果周期效应是乘性的, 则可以采用算子  $\nabla_{12}^2$ 。有时算子  $\nabla_{12}$  本身就足够了, 不必外加差分。对于季节的数据, 可以采用算子  $\nabla_4$  等。

第二步是选定一个特定的模型拟合所分析的时间序列数据。模型识别是 Box-Jenkins 方法中很重要的一环, 比较模型是否合适的一般方法是: 对一般 ARMA 模型体系中的一些特征, 分析其理论特征, 把这种特定模型的理论特征作为鉴别实际模型的标准, 观测实际资料与理论特征的接近程度, 最后根据这种分类比较分析的结果来判定实际模型的类型。

第三步是用时间序列的数据估计模型的参数, 并进行检验, 以判定该模型是否恰当。如不恰当, 则返回第二步, 重新选定模型。

### (1) 模型。

设  $a_t(t=1,2,\cdots,n)$  是服从均值为 0, 方差为  $\sigma^2$  的正态分布的白噪声(可以当作随机误差来理解)序列,  $Y_t$  为等间隔时间  $t$  上的过程值,  $\tilde{Y}_t = Y_t - \mu$  为关于均值  $\mu$  的偏差, 则

$$\tilde{Y}_t = \phi_1 \tilde{Y}_{t-1} + \phi_2 \tilde{Y}_{t-2} + \cdots + \phi_p \tilde{Y}_{t-p} + a_t$$

称为  $p$  阶自回归(AR)过程, 即自回归模型。

如果定义  $p$  阶自回归算子为

$$\varphi_p(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$$

式中,  $B$  为具有  $BY_t = Y_{t-1}$  及  $Ba_t = a_{t-1}$  的后移算子,  $p$  为非季节性自回归模型部分的阶数, 则自回归模型就可简记为

$$\varphi_p(B)\tilde{Y}_t = a_t$$

自回归过程可能是平稳的, 也可能是非平稳的。平稳的必要条件是:  $\varphi_p(B)$  被视为  $B$  的  $p$  阶 AR 多项式时,  $\varphi_p(B) = 0$  的所有根的绝对值都必须大于 1, 也就是所有根都在单位圆外。

自回归模型把过程的偏差  $\tilde{Y}_t$  表示为  $p$  个过去偏差  $\tilde{Y}_{t-1}, \tilde{Y}_{t-2}, \cdots, \tilde{Y}_{t-p}$  的有限加权, 另加一个随机冲击  $a_t$ , 等价于可把  $\tilde{Y}_t$  表示为  $a$  的无限加权。如果使  $\tilde{Y}_t$  线性依赖于有限的  $q$  个  $a$  的过去值, 则得到

$$\tilde{Y}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

式中,  $q$  为非季节性移动平均模型部分的阶数, 称上式为  $q$  阶移动平均(MA)过程。

如果定义  $q$  阶移动平均算子

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

也就是  $\theta_q(B)$  为  $B$  的  $q$  阶 MA 多项式, 则移动平均模型可简记为

$$\tilde{Y}_t = \theta_q(B)a_t$$

当将自回归和移动平均项一同纳入模型时, 就得到

$$\tilde{Y}_t = \varphi_1 \tilde{Y}_{t-1} + \varphi_2 \tilde{Y}_{t-2} + \cdots + \varphi_p \tilde{Y}_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

或

$$\varphi_p(B)\tilde{Y}_t = \theta_q(B)a_t$$

称其为自回归移动平均模型(ARMA)。

在非平稳时, 若存在  $\varphi(B) = 0$  的  $d$  个根在单位圆上, 也即有  $d$  个单位根时, 可以得到

$$\varphi_p(B)(1-B)^d Y_t = \theta_q(B)a_t$$

即

$$\varphi_p(B)\omega_t^d = \theta_q(B)a_t$$

其中

$$\omega_t^d = \Delta^d Y_t$$

式中,  $\Delta$  为差分算子,  $\Delta^d = (1-B)^d$ ;  $d$  为非季节性差分的阶数。

上述过程提供了描述平稳或非平稳时间序列的有效模型, 称为  $(p, d, q)$  阶自回归综合移动平均过程, 记作 ARIMA( $p, d, q$ )。

在时间序列中, 若经过  $s$  个基本时间间隔后, 呈现出相似性, 则称该时间序列有以  $s$  为周期的周期特性, 也称为有季节性趋势, 需用季节模型来拟合;  $s$  称为模型的季节或周期。

假设季节性自回归模型部分的阶数为  $P$ , 季节性移动平均模型部分的阶数为  $Q$ , 季节性差分的阶数为  $D$ ,  $\Phi_p(B^s)$  为  $B$  的  $p$  阶季节性 AR 多项式, 即

$$\Phi_p(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{s^2} - \cdots - \Phi_p B^{sp}$$

$\Theta_Q(B^s)$  为  $B$  的  $Q$  阶季节性 MA 多项式, 即

$$\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{s^2} - \cdots - \Theta_Q B^{sQ}$$

用普通的 ARIMA 模型来模拟含有季节性趋势的时间序列时, 会导致参数过多, 模型过分复杂的问题, 故可采用季节性乘积模型可获取相对简约模型。一般的季节性乘积模型可表示为

$$\varphi_p(B)\Phi_p(B^s)\Delta^s\Delta_d^D Y_t = \theta_q(B)\Theta_Q(B^s)a_t$$

这个模型称为  $(p, d, q) (P, D, Q)$  ARIMA 模型。

我们知道, 用以描述时间序列  $\{Y_t\}$  的 ARIMA 模型

$$\varphi_p(B)Y_t = \theta_q(B)a_t$$

也可用线性滤波运算

$$Y_t = \varphi_p^{-1}(B)\theta_q(B)a_t$$

来表示  $Y_t$  和  $a_t$  之间的关系。这样, 时间序列就表示为一个动态系统的输出, 输入为白噪声, 而其传递函数可以简约地用  $B$  的两个多项式之比来表示。

传递函数 (TF) 模型可以构成较多的模型, 包括作为一个特殊情形的单变量的 ARIMA 模型。假设  $Y_t$  为因变量时间序列, 随意地,  $X_{1t}, X_{2t}, \dots, X_{kt}$  是用于这个模型的预测变量时间序列,  $Z\sigma_t^2$  为  $Z_t$  的预测方差,  $N\sigma_t^2$  为噪声预测值的预测方差, 则一个描述了因变量和预测序列间的关系的 TF 模型可用下式表示:

$$Z_t = f(Y_t)$$

$$\Delta Z_t = \mu + \sum_{i=1}^k \frac{\text{Num}_i}{\text{Den}_i} \Delta_i B^{b_i} f_i(X_{it}) = \frac{MA}{AR} a_t$$

式中, Num 为分子; Den 为分母;  $\Delta$  为差分算子,  $\Delta = (1-B)^d(1-B^s)^D$ 。

单变量的 ARIMA 模型仅仅撤出了 TF 模型中的预测因子; 因此, 它有下列的形式:

$$\Delta Z_t = \mu + \frac{MA}{AR} a_t$$

该模型的主要特征是:

- ① 因变量和预测序列的初始转换,  $f$  及  $f_i$ 。这种转换是任选的且仅当因变量序列值为正数时可用。允许的转换是对数和平方根。这些转换有时被称为方差稳定性转换。
- ② 常数项  $\mu$ 。
- ③ 没有观察到独立且同分布、均值为 0、具有方差  $\sigma^2$  的高斯误差过程  $a_t$ 。
- ④ 移动平均滞后多项式  $MA = \theta_q(B)\Theta_Q(B^s)$  和自回归滞后多项式  $AR = \varphi_p(B)\Phi_P(B^s)$ 。
- ⑤ 差分/滞后算子  $\Delta$  和  $\Delta_i$ 。
- ⑥ 延迟项  $B^{b_i}$ , 其中  $b_i$  是延迟的阶数。
- ⑦ 假定给出了预测因子, 其分子和分母滞后多项式为

$$\text{Num}_i = (\omega_{i0} - \omega_{i1}B - \dots - \omega_{iu}B^u)(\Omega_{i0} - \Omega_{i1}B^s - \dots - \Omega_{iv}B^{vs})B^b$$

以及

$$\text{Ben}_i = (1 - \delta_{i1}B - \dots - \delta_{ir}B^r)(1 - \Delta_{i1}B^s - \dots)$$

- ⑧ “噪声” 序列

$$N_t = \Delta Z_t - \mu - \sum_{i=1}^k \frac{\text{Num}_i}{\text{Den}_i} \Delta_i B^{b_i} X_{it}$$

被假定为均值为 0 的平稳的 ARIMA 过程。

(2) ARIMA/TF 的初始化。

这只需对使用于最优化目标函数的非线性优化算法稍作修改即可。修改考虑“可容许”参数方面的约束。可容许约束要求 AR 的平方根和 MA 多项式超出单位圆, 并且每个预测因

子的分母多项式参数的总和为非 0。最小化算法需要一个初始值来开始其迭代搜索。所有分子和分母多项式参数用 0 来初始化,但分子多项式中 0 次幂的系数除外,它用相应回归系数来初始化。

① ARIMA 参数的初始化。

假定序列  $Y_t$  服从均值为 0 的 ARIMA(p,q)(P,Q), 即

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

在下文中,  $c_I$  和  $\rho_I$  分别表示  $Y_t$  第  $I$  步滞后自协方差和自相关,  $\hat{c}_I$  和  $\hat{\rho}_I$  表示它们的估计。

滞后  $I$  自协方差定义为

$$c_I = E[(Y_t - \mu)(Y_{t+I} - \mu)]$$

滞后  $I$  自相关函数(ACF)定义为

$$\rho_I = \frac{E[(Y_t - \mu)(Y_{t+I} - \mu)]}{\sqrt{E[(Y_t - \mu)^2]E[(Y_{t+I} - \mu)^2]}} = \frac{E[(Y_t - \mu)(Y_{t+I} - \mu)]}{\sigma_Y^2}$$

对自相关函数的估计,统计学家提出了许多方法,也讨论过这些估计的性质,认为  $I$  步滞后自相关  $\rho_I$  最令人满意的估计为

$$\hat{\rho}_I = \frac{c_I}{c_0}$$

式中

$$\hat{c}_I = \frac{1}{N} \sum_{t=1}^{N-I} (Y_t - \bar{Y})(Y_{t+I} - \bar{Y})$$

是自协方差  $c_I$  的估计。

- 非季节性 AR 模型参数。对于 AR 参数的初始值,使用 Box、Jenkins 和 Reinsel(1994)提出的估计方法,其估计为

$$\hat{\phi}'_1, \cdots, \hat{\phi}'_{p+q}$$

- 非季节性 MA 模型参数。

令

$$w_t = Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

互协方差

$$\lambda_I = E(w_{t+I} a_t) = E((a_{t+I} - \theta_1 a_{t+I-1} - \cdots - \theta_q a_{t+I-q}) a_t) = \begin{cases} \sigma_a^2 & I=0 \\ -\theta_1 \sigma_a^2 & I=1 \\ \vdots & \vdots \\ -\theta_q \sigma_a^2 & I=q \\ 0 & I>q \end{cases}$$

假定 AR(p+q) 采用下式可近似估计  $Y_t$ :

$$Y_t - \phi'_1 Y_{t-1} - \cdots - \phi'_p Y_{t-p} - \phi'_{p+1} Y_{t-p-1} - \cdots - \phi'_{p+q} Y_{t-p-q} = a_t$$

本模型的 AR 参数按照上述方法来估计并被表示为

$$\hat{\phi}'_1, \cdots, \hat{\phi}'_{p+q}$$

从而,  $\lambda_I$  可用



$$\begin{aligned}\lambda_I &\approx E[(Y_{t+1} - \varphi_1 Y_{t+1-1} - \cdots - \varphi_p Y_{t+1-p})(Y_t - \varphi_1' Y_{t-1} - \cdots - \varphi_p' Y_{t-p-q})] \\ &= \left( \rho_I - \sum_{j=1}^{p+q} \varphi_j \rho_{I+j} - \sum_{i=1}^p \varphi_i \rho_{I-i} + \sum_{i=1}^p \sum_{j=1}^{p+q} \varphi_i \varphi_j \rho_{I+j-i} \right) c_0\end{aligned}$$

来估计, 并且误差方差  $\hat{\sigma}^2$  下式来近似估计:

$$\hat{\sigma}_a^2 = \text{VAR} \left( -\sum_{j=0}^{p+q} \varphi_j' Y_{t-j} \right) = \sum_{i=0}^{p+q} \sum_{j=0}^{p+q} \varphi_i' \varphi_j' c_{i-j} = c_0 \sum_{i=0}^{p+q} \sum_{j=0}^{p+q} \varphi_i' \varphi_j' \rho_{i-j}$$

并且有  $\hat{\varphi}_0' = -1$ 。

于是, MA 模型参数用  $\theta_I = -\lambda_I / \sigma_a^2$  来逼近, 并用

$$\hat{\theta}_I = -\hat{\lambda}_I / \hat{\sigma}_a^2 = \frac{\rho_I - \sum_{j=1}^{p+q} \hat{\varphi}_j \rho_{I+j} - \sum_{i=1}^p \hat{\varphi}_i \rho_{I-i} + \sum_{i=1}^p \sum_{j=1}^{p+q} \hat{\varphi}_i \hat{\varphi}_j \rho_{I+j-i}}{\sum_{i=0}^p \sum_{j=0}^{p+q} \hat{\varphi}_i \hat{\varphi}_j \rho_{i-j}}$$

来估计。

因此,  $\hat{\theta}_I$  可以用  $\hat{\varphi}_j$ 、 $\hat{\varphi}_i$  以及  $\{\hat{\rho}_I\}_{I=1}^{p+2q}$  来计算。在此过程中, 只使用  $\{\hat{\rho}_I\}_{I=1}^{p+2q}$  且所有其他参数被设为 0。

- 季节性的 AR、MA 模型参数。对于季节性的 AR 和 MA 模型, 在以上等式中使用季节性滞后处的自相关。

### (3) ARIMA/TF 的估计和预测。

有两种预测算法可用: 条件最小二乘法 (CLS) 和极大似然法 (ML)。这两种算法只有一个区别: 它们预测噪声的过程不同。在预测计算中的一般步骤如下:

- ① 在整个历史时期计算噪声过程  $N_t$ 。
- ② 预测噪声过程  $N_t$  直到预测基准线。在历史时期期间, 这是领先一步预测, 在此之后是领先多步预测。在这步中, CLS 和 ELS 预测方法的差别显现出来, 还将计算噪声预测法的预测方差。

③ 通过首先加回噪声预测的常数项的贡献和传递函数的输入, 然后整合并返回转换的结果来获取最终的预测结果。

令  $\hat{N}_t(k)$  和  $\sigma_t^2(k)$  分别为  $k$  步预测值和预测方差。

- 条件最小二乘法 (CLS)。

假定  $t < 0$  时,  $N_t = 0$ , 有

$$\begin{aligned}\hat{N}_t(k) &= E(N_{t+k} | N_t, N_{t-1}, \cdots) \\ \sigma_t^2(k) &= \sigma^2 \sum_{j=0}^{k-1} \psi_j^2\end{aligned}$$

式中,  $\psi_j$  为 MA( $\Delta$ AR) 的幂级数展开系数。

使  $S = \sum [N_t - \hat{N}_t(I)]^2$  最小化。缺失值用  $N_t$  的预测值来插补。

- 极大似然法 (ML) (Brockwell and Davis, 1991)。

$$\hat{N}_t(k) = E(N_{t+k} | N_t, N_{t-1}, \cdots, N_1)$$

$\{N_t - \hat{N}_t(I)\}_{t=1}^t$  的极大似然比, 即

$$L = -\ln(S / n) - (1 / n) \sum_{j=1}^n \ln(\eta_j)$$

式中,  $S = \sum [N_t - \hat{N}_t(I)]^2 / \eta_t$ 。 $\sigma_i^2 = \sigma^2 \eta_t$  是提前一步预测方差。

当出现缺失值时, 使用卡尔曼滤波器计算  $\hat{N}_t(k)$ 。

3. 诊断统计量

ARIMA/TF 诊断统计量基于噪声过程的残差,  $R(t) = N(t) - \hat{N}(t)$ 。

Ljung-Box 统计量为

$$Q(K) = n(n + 2) \sum_{k=1}^K r_k^2 / (n - k)$$

式中,  $r_k$  是第  $k$  步滞后 ACF 的残差。

$Q(K)$  的渐近分布为  $\chi^2[(K - m)]$ ,  $m$  是除了常数项和预测变量有关参数外的参数的数量。

4. 误差方差

在两个模型中误差方差为

$$\hat{\sigma}^2 = S / (n - k)$$

式中,  $n$  为非 0 残差的数量;  $k$  为参数的数量(不包括误差方差)。

5. 拟合优度统计量

拟合优度统计量是根据原始序列  $Y(t)$  计算得到的。设  $k =$  模型中参数的数量,  $n =$  非缺失值残差的数量。

均方差为

$$MSE = \frac{\sum [Y(t) - \hat{Y}(t)]^2}{n - k}$$

平均绝对误差百分比为

$$MAPE = \frac{100}{n} \sum |Y(t) - \hat{Y}(t) / Y(t)|$$

最大绝对误差百分比为

$$MAXAPE = 100 \max(|Y(t) - \hat{Y}(t) / Y(t)|)$$

平均绝对误差为

$$MAE = \frac{1}{n} \sum |Y(t) - \hat{Y}(t)|$$

最大绝对误差为

$$MAXAE = (\sum |Y(t) - \hat{Y}(t)|)$$

标准贝叶斯信息准则(BIC)为

$$\text{标准的 BIC} = \ln(\text{MSE}) + K \frac{\ln(n)}{n}$$

$R$  方为

$$\text{MSE} = \frac{\sum [Y(t) - \hat{Y}(t)]^2}{\sum (Y(t) - \bar{Y})^2}$$

平稳  $R^2$  为

$$R_s^2 = 1 - \frac{\sum (Z(t) - \hat{Z}(t))^2}{\sum (\Delta Z(t) - \overline{\Delta Z})^2}$$

式中,  $Z(t) - \hat{Z}(t)$  和  $\Delta Z(t) - \overline{\Delta Z}$  中的所有项的和是不可缺少的。

$\overline{\Delta Z}$  是差分转换序列的简单平均模型, 它等价于单变量基准模型  $\text{ARIMA}(0, d, 0)$   $(0, D, 0)$ 。

对于当前正在考虑的指数平滑模型, 使用差分的阶数(如果有的话, 则相当于其等价于  $\text{ARIMA}$  模型), 即

$$d = \begin{cases} 2 & \text{Brown, Holt} \\ 1 & \text{其他} \end{cases}, \quad D = \begin{cases} 0 & s = 0 \\ 1 & s > 1 \end{cases}$$

注意: 平稳和常规的  $R^2$  在  $(-\infty, 1]$  范围内取负值。负的  $R^2$  意味着考虑的模型要比基准模型更差;  $R^2$  为 0 意味着考虑的模型与基准模型差不多; 正的  $R^2$  意味着考虑的模型要比基准模型更好。

## 6. 专家建模器

专家建模器可用来对单变量和多变量时间序列进行自动建模分析。

(1) 变量时间序列。

对于单变量时间序列, 用户可以让专家建模器为其从以下模型中选择一个模型建模:

- ① 默认状态下, 选择所有模型。
- ② 只是指数平滑模型。
- ③ 只是  $\text{ARIMA}$  模型。

(2) 多变量序列。

在多变量情况下, 用户可以让专家建模器为他们从下面的模型中选择一种模型建模:

① 默认状态下, 选择所有模型。注意如果多变量  $\text{ARIMA}$  专家模型放弃所有预测变量并用单变量  $\text{ARIMA}$  专家模型结束, 则这个单变量  $\text{ARIMA}$  专家模型将与先前的专家指数平滑模型作比较, 而且专家建模器将决定所有模型中哪个模型更好。

② 只是  $\text{ARIMA}$  模型。

(3) 使用专家建模器自动选择模型。

① 选择所有模型。在这种情况下, 计算指数平滑和  $\text{ARIMA}$  专家模型, 并且选择具有更小的标准化  $\text{BIC}$ (贝叶斯信息标准)的模型。对于  $n < \max(20, 3s)$  的短的时间序列, 使用指数平滑专家模型。

② 选择指数平滑专家模型。选择本模型, 对于  $1 < n \leq 10$  的短时间序列, 专家建模器将自动拟合简单  $\text{ES}$  指数平滑模型; 对于其他时间序列, 专家建模器自动按如图 17-9 所示的流程选择模型。

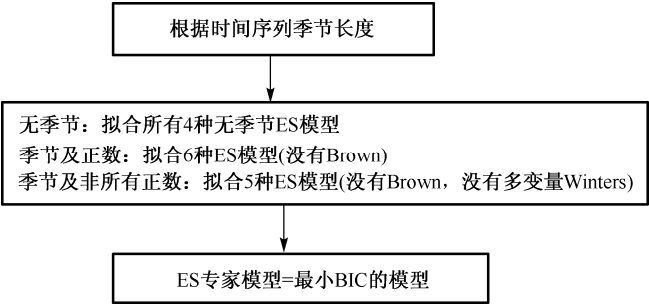


图 17-9 指数平滑专家模型自动建模流程

③ ARIMA 专家模型。对于  $n < 10$  的短时间序列, 拟合有常数项的 AR(1), 如果  $10 < n < 3s$ , 则设  $s = 1$ , 构建非季节模型; 对于其他情形, 专家建模器自动按如图 17-10 所示的流程选择模型。

④ 选择 TF 模型。选择本模型, 对于  $n < \max(20, 3s)$  的短时间序列, 专家建模器将自动拟合单变量专家模型; 对于其他时间序列, 专家建模器自动按如图 17-11 所示的流程选择模型。

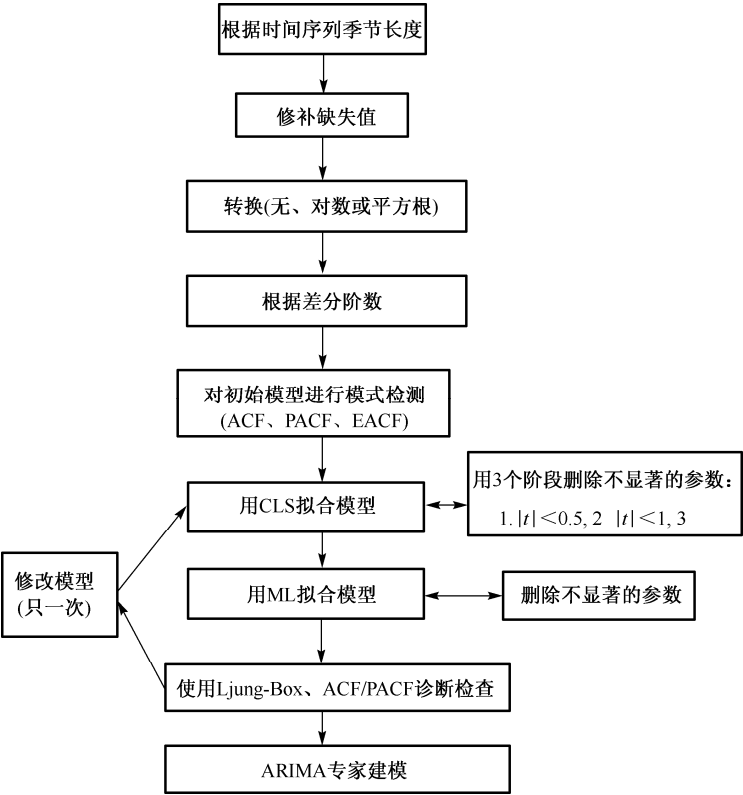


图 17-10 ARIMA 专家建模流程

7. 在时间序列分析中对异常值的检测

异常值也称离群值。观察序列可以受到所谓异常值的污染。这些异常值可以改变未被污染的序列的平均水平。异常值检测的目的是寻找是否有异常值以及异常值的位置、类型和大小。

(1) 异常值的定义。

自动建模过程(TSMODEL)中考虑 7 种类型的异常值: 加性异常值(AO)、创新性异常值

(IO)、水平移位 (LS) 异常值、临时 (或短暂) 变更 (TC) 异常值、季节性加性 (SA) 异常值、局部趋势 (LT) 异常值以及加性异常值小块 (AOP)。

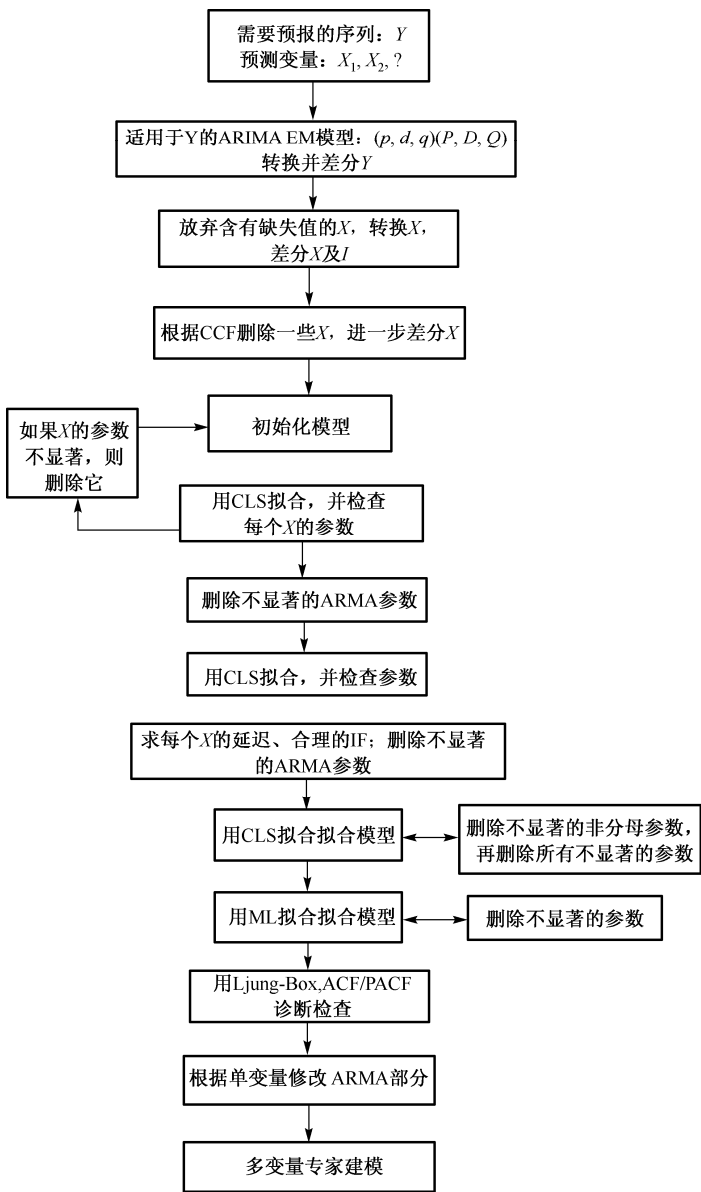


图 17-11 TF 专家建模流程

假设  $U(t)$  或  $U_t$  为未被污染的序列, 无异常值约束。假定其为单变量 ARIMA 或传递函数模型。

① AO。假定一个 AO 异常值发生在时间  $t = T$ , 观察序列可表示为

$$Y(t) = U(t) + wI_T(t)$$

式中,  $I_T(t) = \begin{cases} 0 & t \neq T \\ 1 & t = T \end{cases}$  是一个脉冲函数;  $w$  是由异常值引起的真正的  $U(t)$  的偏差。造成这种异常值的干扰, 只影响干扰发生的那一时刻  $T$  上的序列值, 而不影响该时刻以后的序列值。

② IO。假定一个 IO 异常值发生在时间  $t = T$ ，观察序列可表示为

$$Y(t) = \mu(t) + \frac{\theta(B)}{\Delta\varphi(B)}[a(t) + wI_T(t)]$$

③ LS。假定一个 LS 异常值发生在时间  $t = T$ ，观察序列可表示为

$$Y(t) = U(t) + wS_T(t)$$

式中， $S_T(t) = \frac{1}{1-B}I_T(t) = \begin{cases} 0 & t < T \\ 1 & t \geq T \end{cases}$  是一个阶梯函数。

④ TC。假定一个 TC 异常值发生在时间  $t = T$ ，观察序列可表示为

$$Y(t) = U(t) + wD_T(t)$$

式中， $D_T(t) = \frac{1}{1-\delta B}I_T(t)$ ,  $(0 < \delta < 1)$ ，是一个阻尼函数。

⑤ SA。假定一个 SA 异常值发生在时间  $t = T$ ，观察序列可表示为

$$Y(t) = U(t) + wSS_T(t)$$

式中， $SS_T(t) = \frac{1}{1-B^s}I_T(t) = \begin{cases} 1 & t = t + ks, k \geq 0 \\ 0 & \text{o.w.} \end{cases}$  是一个阶梯季节性脉冲函数。

⑥ LT。假定一个 LT 异常值发生在时间  $t = T$ ，观察序列可表示为

$$Y(t) = U(t) + wT_T(t)$$

式中， $T_T(t) = \frac{1}{(1-B)^2}I_T(t) = \begin{cases} t+1-T & t \geq T \\ 0 & \text{o.w.} \end{cases}$  是一个本地化趋势函数。

⑦ AOP。一个加性异常值补丁是两个或多个连续的 AO 异常值群。一个加性异常值补丁可用其开始的时间和长度来描述。假定在时间  $t = T$  有一个长度为  $k$  的 AO 异常值小块(patch)，观察序列可表示为

$$Y(t) = U(t) + \sum_{i=1}^k w_i I_{T-1+i}(t)$$

由于掩蔽效应，当逐个搜查异常值时，检测 AO 异常值补丁是非常困难的。这就是为什么把 AO 异常值补丁从单个 AO 中分离出来作为单独一类的原因。对于 AO 异常值补丁，程序对所有的补丁一起搜寻。

对于时间  $t = T$  的类型 O 的异常值(AO Patch 除外)，可汇总如下：

$$Y(t) = \mu(t) + wL_O(B)I_T(t) + \frac{\theta(B)}{\Delta\varphi(B)}a(t)$$

式中，

$$L_O(B) = \begin{cases} 1 & O = AO \\ 1/\Delta\pi(B) & O = IO \\ 1/(1-B) & O = LS \\ 1/(1-\delta B) & O = TC \\ 1/(1-B^s) & O = SA \\ 1/(1-B)^2 & O = LT \end{cases}$$

以及  $\pi B = \varphi(B) / \theta(B)$ 。合并异常值的一般模型因而可以写成下式：

$$Y(t) = \mu(t) + \sum_{k=1}^M w_k L_{Ok}(B) I_{Tk}(t) + \frac{\theta(B)}{\Delta\varphi(B)} a(t)$$

式中,  $M$  是异常值的数量。

在实践中, 这些异常值类型的组合可出现在研究的序列中。

## (2) 估计异常值的影响。

假如模型和模型参数已知, 还假定异常值的类型和位置已知, 异常值大小的估计和检验统计量如下。

在这部分的结果只使用于异常值检测过程的中间步骤。异常值的最终估计来自于模型合并的所有异常值, 所有参数被连带估计。

图 17-12 所示的流程图说明了如何自动开展异常值的检测工作。设  $M$  为异常值的总数,  $Nadj$  为时间序列被调整的异常值的数量。在程序开始,  $M = 0$  及  $Nadj = 0$ 。

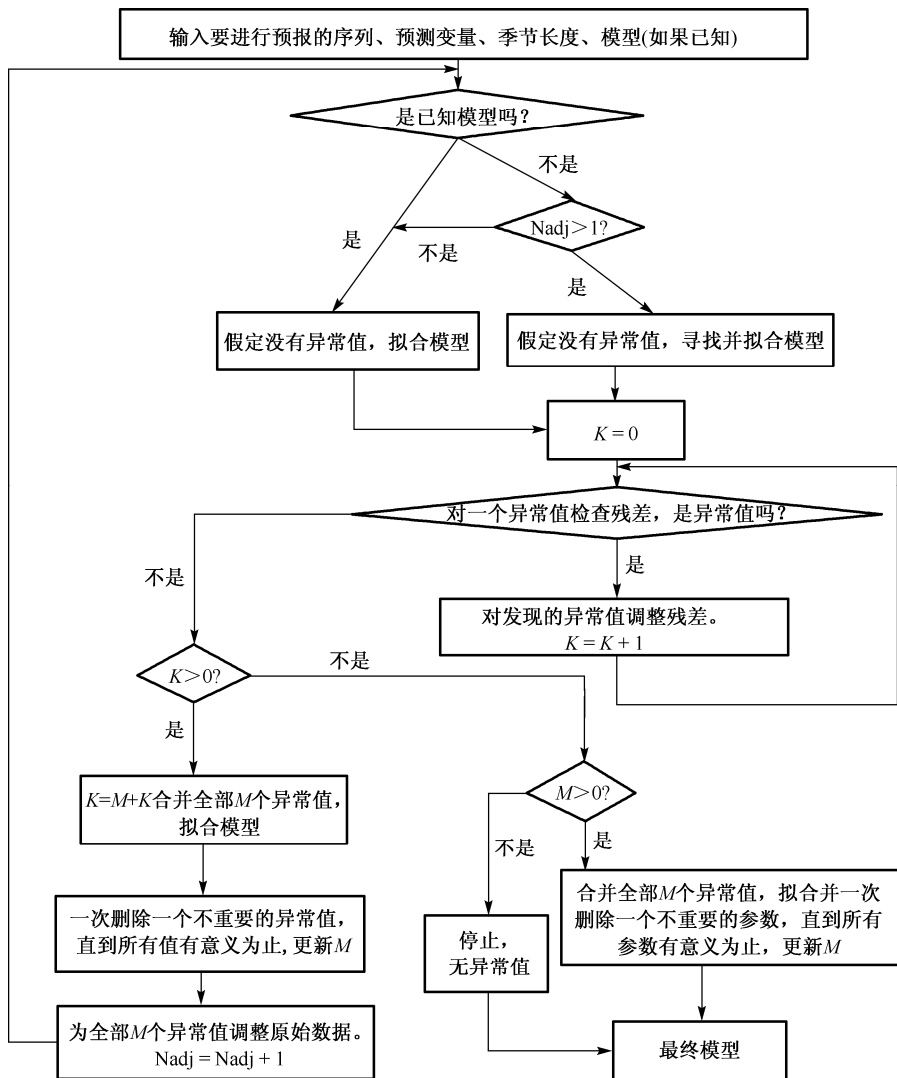


图 17-12 异常值检测过程

### 17.3.2 选择分析变量

在【时间序列建模】提示框中,单击【确定】按钮,关闭提示框,打开如图 17-13 所示的【时间序列建模器】对话框【变量】选项卡。

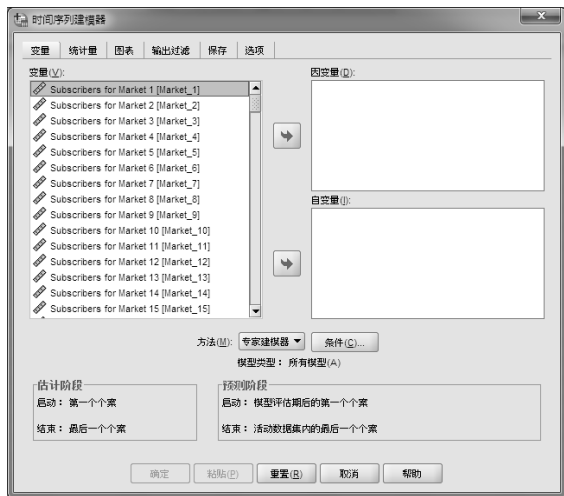


图 17-13 【时间序列建模器】对话框【变量】选项卡

(1) 定义因变量和自变量。

在变量框中选定一个或多个变量送到【因变量】框中,作为因变量。

如果建模需要,则也可在变量框中选定一个或多个变量送到【自变量】框中,作为自变量。

因变量和自变量都应是数值变量,它们都会被认为是时间序列,也就是说,每个样品代表了一个时间点,连续的样品通过一个恒定的时间间隔分开。

(2) 【估计阶段】(估计期)栏显示了估计期的起始和结束位置。估计期即用来估计模型的样品集。默认的是从第一个观测到最后一个观测。如果要改变估计期,使用数据窗【数据】菜单中的【选择个案】功能。在该功能的对话框中选择【基于时间或个案全距】并单击【范围】按钮,在二级对话框中设定估计期。改变了的估计期将显示在如图 17-13 所示对话框的【估计阶段】栏中。

(3) 【预测阶段】(预测期)栏显示了预测期的起始和结束位置。默认的是从估计期结束后的第一个观测开始,到实际数据集的最后一个观测为止。可以在时间序列模型的主对话框【选项】卡中改变预测期,改变后该栏显示新设定的预测期。

例如,时间序列是从 1999 年 1 月开始到 2003 年 12 月结束的 4 年数据。这也是默认估计期的范围。如果定义估计期从 2000 年 1 月开始到 2002 年 12 月止。则默认的预测期从 2003 年 1 月起,到 2003 年 12 月止。

(4) 确定建模方法。

【方法】下拉列表中共有 3 个选项:【专家建模器】、【指数平滑法】和【ARIMA】(自回归综合移动平均法)。其中【专家建模器】是系统默认的建模方法。

(5) 设定模型条件。

单击【条件】按钮,可在弹出的对话框中设定各种建模方法下的模型类型及因变量的转换方式等。



17.3.2.1 专家建模器

【专家建模器】会自动地为每个因变量序列找到最佳拟合模型。

在主对话框【方法】下拉列表中选择【专家建模器】，单击【条件】按钮，打开如图 17-14 所示的【时间序列建模器：专家建模器条件】对话框。

1. 【模型】选项卡

(1) 在【模型类型】栏中选择模型的类型，有 3 个有效选项：在这里选中的模型将显示在主对话框的【模型类型】栏中。

①【所有模型】。系统默认选项。同时考虑指数平滑法和 ARIMA 自回归综合移动平均法，程序会自动识别用哪个作为拟合时间序列的最佳模型。

②【仅限指数平滑模型】。只对时间序列采用指数平滑法进行估计。

③【仅限 ARIMA 模型】，只对时间序列采用自回归综合移动平均法进行估计。

(2)【专家建模器考虑季节性模型】选项。只有当前数据文件已定义周期时才有效。选择该项，【专家建模器】同时考虑季节性和非季节性模式；如果未选择该选项，【专家建模器】只考虑非季节性模式。

在【当前周期性】后面显示已定义的周期。如果没有定义周期，则显示值【无】。

(2)【事件】栏。选择的任何自变量都被当作事件变量。事件变量值为 1 的样品表明该时期的因变量序列被期望受到事件影响；1 以外的值表明不受影响。

2. 【离群值】选项卡

如图 17-15 所示，该选项卡中可以选择自动检测异常值的类型。

(1)【自动检测离群值】选项。默认状态下不会自动检测异常值。

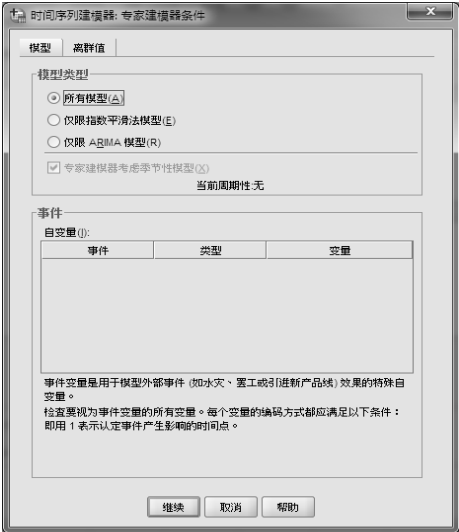


图 17-14 【时间序列建模器：专家建模器条件】对话框【模型】选项卡



图 17-15 【时间序列建模器：专家建模器条件】对话框【离群值】选项卡

(2)【要检测的离群值类型】栏。在此栏中可选择一项或多项异常值类型：

①【加法】。加性异常值(AO)，影响单个观察值的异常值。例如，一个数据编码错误可能

被认为是一个加性异常值。

②【移位水平】。水平移位异常值(LS)。在一个特定的序列点(异常值出现的时间点)开始,所有观察值由其自身值加上一个常量转换而成,该常量等于特定的序列点处的异常值与其真值的偏差。

③【创新的】。在一个特定的序列点开始,异常值中增加了噪声项的作用,称受噪声项影响的异常值为创新性异常值。对平稳序列,创新异常值(Innovational Outlier)只影响少数观察值;但对非平稳序列,创新异常值可能会影响在一个特定的序列点开始的每个观察值。

④【瞬时的】。临时(或短暂)变更的异常值(TC)。

⑤【季节性可加的】。季节性加性异常值(SA)。影响一个特定的观察值和其后由一个或多个周期隔开的所有的观察值。所有这些观察值受到的影响相同。如果从某年开始的各个 1 月份销售额都较高,则周期性加性异常值可能会发生。

⑥【局部趋势】异常值(LT)。在一个特定的序列点开始导致局部线性趋势的异常值。

⑦【可加的修补】。加性异常值小块(AOP)。两个或两个以上连续的加性异常值组。选择该异常值类型会导致对这些序列点以外的单个加性异常值的检测。

单击【继续】按钮,返回图 17-13 所示的主对话框。

17.3.2.2 指数平滑法

仅当只指定了因变量时,在【方法】下拉列表中才能激活【指数平滑法】。

在主对话框的【方法】下拉列表中选择该项,单击【条件】按钮,打开如图 17-16 所示的【时间序列建模器:指数平滑条件】对话框。

1. 【模型类型】栏

【模型类型】栏中有两种模型:非季节性模型、季节性模型。

(1) 在【非季节性】模型中有 4 个选项:

①【简单】模型。适用于无趋势或无季节因素影响的时间序列。唯一的平滑参数是水平。简单指数平滑同具有零阶自回归、一阶差分、一阶移动平均的 ARIMA 模型极其相似,并且没有常量。

②【Holt 线性趋势】模型。适用于有线性趋势和无季节性因素影响的时间序列。其平滑参数为水平和趋势,它不受彼此值的约束。Holt 模型比 Brown 模型更普通,而且计算一个较长的时间序列时要花更长的时间。Holt 指数平滑模型同具有零阶自回归、二阶差分、二阶移动平均的 ARIMA 模型极其相似。

③【Brown 线性趋势】模型。适用于有线性趋势和无季节因素影响的时间序列。其平滑参数为水平和趋势,且假定它们相等。

Brown 模型因此是 Holt 模型的特例。Brown

指数平滑模型同具有零阶自回归、二阶差分、二阶移动平均的 ARIMA 模型很相似,同时第二阶移动平均系数等于第一阶系数一半的平方。

④【阻尼趋势】模型。适用于线性趋势正在消失且无季节因素的时间序列。其平滑参数为



图 17-16 【时间序列建模器:指数平滑条件】对话框

水平、趋势和阻尼趋势 (Damping Trend)。阻尼指数平滑模型与具有一阶自回归、一阶差分、二阶移动平均的 ARIMA 模型很相似。

(2) 在【季节性】模型中有 3 个选项:

①【简单季节性】模型。适用于无趋势和季节因素影响在时间上是常量的时间序列。其平滑参数是水平和季节。简单季节性指数平滑最类似于具有零阶自回归、一阶差分、一阶季节差分、一阶、 $p$  阶及  $p+1$  阶移动平均的 ARIMA 模型。其中,  $p$  是在一个季节间隔中的周期数(如每月数据,  $p=12$ )。

②【Winters 可加性】温特加性模型。适用于有线性趋势且不依赖于序列水平季节性影响的时间序列。其平滑参数是水平、趋势和周期。Winters 加性指数平滑与具有零阶自回归、一阶差分、一阶季节差分及  $p+1$  阶移动平均的 ARIMA 模型极其相似。其中,  $p$  是在一个季节间隔中的周期数(如每月数据,  $p=12$ )。

③【Winters 相乘性】温特积性模型。适用于有线性趋势且依赖于序列水平的周期性影响的时间序列。其平滑参数是水平、趋势和周期。Winters 积性指数平滑模型同任何 ARIMA 模型都不相似。

【当前周期性】后面的整数是当前的数据文件定义的周期。例如, 年周期为 12, 每个样品表示 1 个月。如果没有设定周期, 则显示的值为【无】。季节模型需要设置周期。可以在【数据】菜单的【定义日期】对话框中设置。

## 2. 【因变量转换】栏

该栏定义对时间序列的转换方法。有 3 个选项供选择。在建模之前, 可以对每个因变量实施转换。

①【无】。对时间序列不实施转换。

②【平方根】。对时间序列用平方根转换。

③【自然对数】。对时间序列用自然对数转换。

单击【继续】按钮, 返回图 17-13 所示的对话框。

### 17.3.2.3 ARIMA 自回归综合移动平均模型

在主对话框的【方法】下拉列表中选择【ARIMA】, 单击【条件】按钮打开如图 17-17 所示的【时间序列建模器: ARIMA 条件】对话框【模型】选项卡。



图 17-17 【时间序列建模器: ARIMA 条件】对话框【模型】选项卡

## 1. 在【模型】选项卡中指定自定义模型结构

(1)【ARIMA】阶数栏。

【结构】表的单元格中需要输入非负整数, 以便定义 ARIMA 模型的构成。对自回归和移动平均构成, 值代表最大的阶数。所有比最大阶数值小的阶数值都包含在模型中。例如, 如果指定 2, 则模型包括二阶和一阶。【季节性】列只在当前数据文件已定义周期时有效。

① 非季节性参数的阶数。

- 【自回归(p)】。设置用序列的先前值来预测现值的自回归的阶数。自回归的阶数 2 指定用序列过去 2 个时间周期的值来预测现值。
- 【差分(d)】。指定用于时间序列的差分转换的阶数。当存在趋势时(含有趋势的时间序列典型地是非平稳的, 假设 ARIMA 模型是平稳的), 差分是必要的, 可用来消除趋势的影响。差分的阶数同时间序列趋势的程度相对应, 一阶差分说明线性趋势, 二阶差

分说明二次趋势，等等。

- **【移动平均数(q)】**。指定模型中移动平均的阶数。移动平均阶数定义了用原先值同序列平均值的偏差来预测当前值。例如，一阶和二阶移动平均说明当预测时间序列的当前值时，要考虑最后两个时间周期中的每个值同时间序列的平均值的偏差。

## ② 季节性参数的阶数。

季节性自回归、移动平均以及差分构成同它们在非季节序列对应项含义相同。对于季节性阶数，时间序列当前值受由一个或几个季节周期隔开的先前序列值影响。例如，对于每月数据(季节性周期为 12)，季节阶数 1 意味着当前的序列值受先于当前值 12 个周期的序列值影响。对每月数据，季节阶数 1，等于指定了 12 的非季节阶数。

**【当前周期性】**显示当前数据文件定义的周期，是一个整数。例如，12 代表每年的周期，每个样品表示 1 个月。如果没有设置周期，则显示**【无】**。季节模型需要周期。用户可以在**【数据】**菜单的**【定义日期】**对话框中设置周期。

(2) 在**【转换】**栏中指定各因变量进入模型之前的转换。

① **【无】**。对时间序列不进行任何转换。

② **【平方根】**。对时间序列进行平方根转换。

③ **【自然对数】**。对时间序列进行自然对数转换。

(3) **【在模型中包括常数】**选项。除非确信全部序列值的平均数为 0，否则在模型中应该包含常数项，即应该选择此项。当使用差分时，建议在模型中不要常数项。

## 2. 单击**【离群值】**选项卡，在如图 17-18 所示的对话框中选择对异常值的处理方法

(1) **【不检测离群值或为其建模】**。系统默认异常值不被检测也不进入模型。

(2) **【自动检测离群值】**。自动对异常值进行检测的选项。

在此可以选择要检测的一个或多个异常值的类型。提供选择的异常值类型有**【加法】**、**【移位水平】**、**【创新的】**、**【瞬时的】**、**【季节性可加的】**、**【局部趋势】**、**【可加的修补】**。有关这些选项的说明，请参见 17.3.2.1 节中的相关内容。

(3) **【将特定的时间点作为离群值来建模】**栏。指定特定的时间点作为异常值。每个异常值占**【离群值定义】**表中的一行。在这行的各单元格中输入时间点的数值。

① **【类型】**列的下拉列表中可选择异常值的类型。支持的类型有**【加法】**(系统默认项)、**【移位水平】**、**【创新的】**、**【瞬时的】**、**【季节性可加的】**以及**【局部趋势】**。

② 在**【类型】**前的两列表头根据定义的周期显示不同的内容，如年、月等。在各行相应位置输入表达异常值时间点的数值。注意，如果没有定义日期变量，则**【离群值定义】**表显示单列的观测。为指定一个异常值，输入异常值样品(在**【数据编辑】**窗口中显示的)。

## 3. 在**【转换函数】**选项卡中定义自变量的转换函数

只有在主对话框**【变量】**选项卡中指定了自变量，并在**【方法】**下拉列表中指定了**【ARMIA】**方法，单击**【条件】**按钮打开的对话框中才会有**【转换函数】**选项卡，见图 17-19。

在该选项卡中，对在**【变量】**选项卡中指定的一个或多个自变量指定转换函数。转换函数允许指定使用自变量(预测变量)的过去值来预报因变量的将来值的方法。

(1) **【转换函数的阶数】**栏。

在**【结构】**表格单元格中输入转换函数不同成分的值。所有值必须是非负的整数。对于**【分子】**和**【分母】**，输入的值代表最大的阶数。所有比指定值小且大于 0 的阶数都会包括在模型

中。此外，0 阶始终包含在分子成分中。例如，如果在【分子】后面的单元格中输入“2”，那么模型包括二阶、一阶和 0 阶。如果在【分母】后面的单元格中输入“3”，那么模型包括三阶、二阶和一阶。如果【季节性】列没有被激活，那是因为在当前工作的数据文件中没有定义周期。



图 17-18 【时间序列建模器：ARIMA 条件】对话框【离群值】选项卡

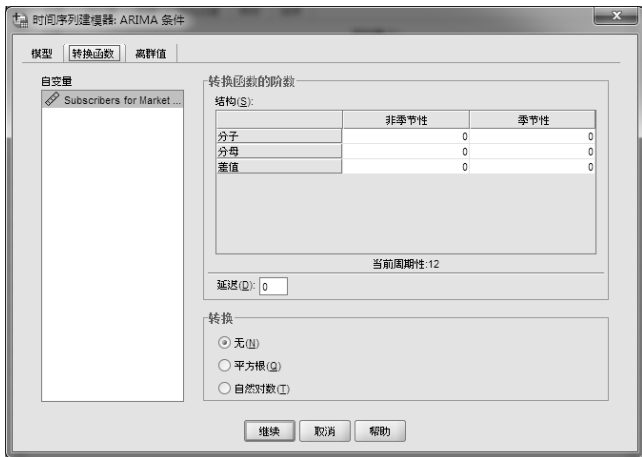


图 17-19 【时间序列建模器：ARIMA 条件】对话框【转换函数】选项卡

①【分子】。指定转换函数分子的阶数。用选择的自变量(预测变量)序列中指定阶数的先前值去预测因变量的当前值。例如，一阶的分子指定用过去一个时间周期的自变量序列的值，以及自变量序列的现值，预测各个因变量序列的现值。

②【分母】。指定转换函数分母的阶数。指定选择的自变量(预测变量)序列的指定阶数的先前值与时间序列均数的偏差来预测因变量的当前值。例如，一阶的分母指定在预测各因变量序列的现值时，要考虑自变量序列过去一个时间周期的平均值的偏差。

③【差值】(差分)。指定在估计模型前用于选择的自变量(预测变量)序列差分的阶数。当存在趋势及要消除它的影响时，需要用差分。

④【季节性】阶数。季节性的分子、分母和差分成分同非季节性的分子、分母和差分成分发挥同样的作用。对季节性阶数，当前序列值受到由一个或多个季节周期分开的先前序列值的影响。例如，对每月的数据(季节周期 12)而言，季节阶数 1 意味着当前序列值受到先于当前值 12 个周期的序列值的影响。

(2)【当前周期性】栏。指出在工作的数据集中定义过的当前周期(如果有的话)。

(3)【延迟】框。通过指定一个间隔数设置延迟，促使自变量的影响相应延后。例如，如果设置延迟为“5”，则在时间  $t$  的自变量的值不影响预报，而  $t+5$  后的自变量的值对预报有影响。

(4)【转换】栏。选择对自变量作转换的方法。

- 【无】。不作转换。它是默认选项。
- 【平方根】。用平方根转换。
- 【自然对数】。用自然对数转换。

### 17.3.3 选择统计量

在主对话框中单击【统计量】选项卡，得到图 17-20 所示的【时间序列建模器】对话框【统计量】选项卡。选择输出建模结果的选项。



图 17-20 【时间序列建模器】对话框【统计量】选项卡

(1) 【按模型显示拟合(优度的)度量、Ljung-Box 统计量和离群值的数量】。显示的表格中包括选择的拟合优度的测度、Ljung-Box 值以及各模型的异常值数。

(2) 【拟合度量】栏。选择拟合测度。

① 【平稳的  $R^2$ 】。将模型的平稳部分和简单平均模型进行比较。当有趋势或季节模式时，本测度比普通  $R^2$  更好。平稳  $R^2$  方值范围是负无穷大到 1。负值意味着在考虑中的模型比基准模型更糟，正值意味着在考虑中的模型比基准模型更好。

② 【 $R^2$ 】。即  $R^2$ ，由模型解释的时间序列中的总变异比例的估计。当时间序列是平稳序列时，本测度极有用。 $R^2$  值的范围是负无穷大到 1。负值表示考虑中的模型不如基准模型，正值表示优于基准模型。

③ 【均方根误差】。因变量序列同其模型预测值间差异程度的测度，用与因变量序列相同的单位来表示。

④ 【平均绝对误差百分比】。因变量序列同它的模型预测值间差异程度的测度。由于它不依赖于使用的单位，因此它能用来比较不同单位的序列。

⑤ 【平均绝对误差】(MAE)。因变量序列同它的预测模型水平间差异程度的测度。MAE 采用原先序列的单位。

⑥ 【最大绝对误差百分比】。用百分比表示的最大预测误差。对于预测设想一个最坏的结果时，本测度很有用。

⑦ 【最大绝对误差】。最大预测误差，用因变量序列的相同单位来表示。像最大绝对误差百分比一样，对预测设想一个最坏的结果时，本测度很有用。最大绝对误差和最大绝对误差百分比可以发生在不同的序列点。例如，当一个大的序列值的绝对误差比一个小序列值的绝对误差稍大时，最大绝对误差将发生在较大的序列值处，最大绝对百分比误差将发生在较小的序列值处。

⑧ 【标准化的 BIC】。标准化贝叶斯信息准则。它是试图说明模型复杂性对模型整体拟合的综合测度，是建立在均方误差基础上的得分，并包括在模型中参数的数量和序列在长度上的损失。这个损失降低了有更多参数的模型的优势，从而使统计量更容易对同一序列的不同模型进行比较。

(3) 【比较模型的统计量】栏。选择比较模型的统计量，控制如何显示包括所有估计模型的统计计算的表格。每个选项会产生一张单独的表。

① 【拟合优度】。统计量，产生平稳  $R^2$ 、 $R^2$ 、均方误差的平方根、均数绝对百分比误差、均数绝对误差、最大绝对百分比误差以及标准化贝叶斯信息准则的摘要统计表和百分比表。

② 【残差自相关函数(ACF)】。产生所有估计模型残差自相关的摘要统计表和百分比表。

③ 【残差部分(应为偏)自相关函数(PACF)】。产生所有估计模型残差偏自相关的摘要统计表和百分比表。

(4) 【个别模型的统计量】栏。选择包括每个估计模型详情的表格。每个选项产生一张单独的表。

① 【参数估计】。为每个估计模型显示一张参数估计表。为指数平滑模型和 ARIMA 模型显示单独的表。如果异常值存在，也会显示在一张单独的表中。

② 【残差自相关函数(ACF)】。为每个估计模型显示一张滞后的残差自相关表，包括自相关的置信区间。

③ 【残差部分(应为偏)自相关函数】(PACF)。为每个估计模型显示一张滞后的残差偏自相关表，包括偏自相关的置信区间。

(5) 【显示预测值】选项。为每个估计模型显示一张模型预测和置信区间表。预测期可在【选项】卡中设置。

### 17.3.4 图表

单击【图表】选项卡，打开如图 17-21 所示的【时间序列建模器】对话框【图表】选项卡。在该选项卡中选择建模结果图。

(1) 【模型比较图】栏。选择表现模型拟合程度的图形，其中前 8 个选项对应【统计量】选项卡中【拟合度量】栏中的 8 个统计量。每个选项单独产生一个图形，可以多项同时选择。可供选择的选项有【平稳的 R 方】、【R 方】、【均方根误差】、【平均绝对误差百分比】、【平均绝对误差】、【最大绝对误差百分比】、【最大绝对误差】、【标准化的 BIC】。统计量解释详见 17.3.3 节。此外还有两个选项：

① 【残差自相关函数(ACF)】。在图形中包含残差自相关函数。

② 【残差部分(应为偏)自相关函数(PACF)】。在图形中包含残差偏自相关函数。

(2) 【单个模型图】栏中的选项是针对单个模型的。

- 【序列】。对每个估计模型产生预测值图。每个图包含的内容可以选择下列选项：
- 【观察值】。在图形中包含因变量序列的观察值。
- 【预测值】。图中显示预测期中的模型预测值。
- 【拟合值】。在图形中包含估计期的模型预测值。



图 17-21 【时间序列建模器】对话框【图表】选项卡

- **【预测值的置信区间】**。在图形中包含预测值的置信区间。
- **【拟合值的置信区间】**。在图形中包含估计值的置信区间。
- (3) **【残差自相关函数】(ACF)**，为各估计模型显示残差自相关图。
- (4) **【残差部分(应为偏)自相关函数(PACF)】**。显示各估计模型残差偏自相关图。

17.3.5 输出项目的过滤

单击**【输出过滤】**选项卡，打开如图 17-22 所示的**【输出过滤】**选项卡，对估计模型子集的表和图的输出进行限制。

- (1) **【在输出包含所有的模型】**。系统默认输出包含所有的估计模型。
- (2) **【基于拟合优度过滤模型】**。根据拟合优度限制模型的输出。

- ① **【最佳拟合模型】**。输出最佳拟合模型。选择此项还要输入限制参数：
  - **【模型的固定数量】**。指定显示  $n$  个最佳拟合模型。如果  $n$  大于估计模型的数量，则显示所有的模型。
  - **【占模型总数的百分比】**。显示模型中拟合优度最高的前  $n\%$  个模型。

- ② **【最差拟合模型】**。输出最差拟合模型。选择此项还要输入限制参数：
  - **【模型的固定数量】**。指定显示  $n$  个最差拟合模型的结果。如果指定的数量超过估计模型的数量，则显示所有的模型。
  - **【占模型总数的百分比】**。显示模型中拟合优度最低的最后  $n\%$  个模型。

- ③ **【拟合优度】**的测度。选择以上两种过滤方式，还需要指定用于过滤模型的拟合优度测度。下拉列表中共有 8 个选项，系统默认为**【平稳的 R 方】**。它们分别是**【平稳的 R 方】**、**【R 方】**、**【均方根误差】**、**【平均绝对误差百分比】**、**【平均绝对误差】**、**【最大绝对误差百分比】**、**【最大绝对误差】**、**【标准化的 BIC】**。有关这些选项的说明，详见 17.3.3 节中的相关内容。

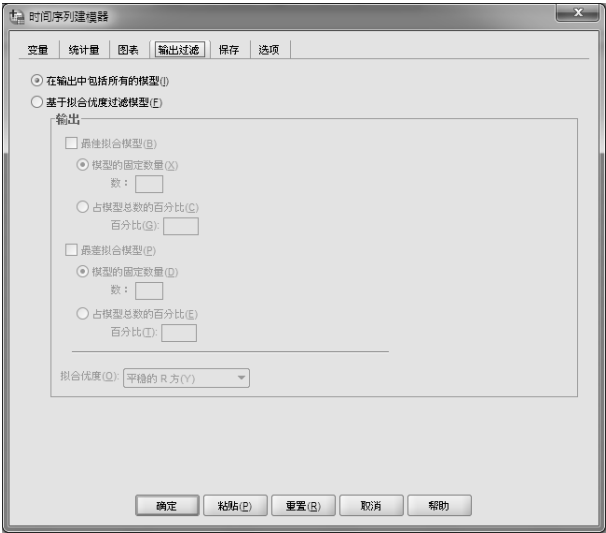


图 17-22 **【时间序列建模器】**对话框**【输出过滤】**选项卡

17.3.6 保存新变量

单击**【保存】**选项卡，打开如图 17-23 所示的**【时间序列建模器】**对话框**【保存】**选项卡，指定要保存在当前工作的数据文件的新变量和保存到外部文件的选项。

仅当在**【输出过滤】**选项卡中选择了**【最佳拟合模型】**或**【最差拟合模型】**选项，并设定完毕后，**【保存】**选项卡才被激活。

1. 保存新变量

在**【保存变量】**栏中，可以在当前工作的数据文件中存储模型预测值、置信区间以及残差的新变量。每个因变量序列建立与自身有关的新变量，每个新变量包含估计和预测期的值。如果预测期超过因变量序列的长度，则增加新样品。通过为以下各个统计量选择相关的**【保存】**选项，选择存储新变量。在系统默认情况下，不保存新变量。



- ①【预测值】。存储模型预测值。
- ②【置信区间的下限】。存储预测值置信区间的下限。
- ③【置信区间的上限】。存储预测值置信区间的上限。
- ④【噪声残差】。存储模型残差。如果对因变量进行了转换，如用自然对数进行转换，存储的是转换序列的残差。
- ⑤【变量名的前缀】。指定新变量名的前缀，或留下默认的前缀。变量名由前缀、与因变量有关的名字以及模型标识符组成。如果必须避开变量名的冲突，则变量名用扩展名。前缀必须符合有效变量名的命名规则。

2. 对输出模型文件进行命名

对所有估计模型的模型说明被输出到 XML 格式的指定文件中。存储的模型可以对更多的数据，使用【应用模型】程序，获取新的预测。

17.3.7 建模的其他选项

单击【选项】选项卡，打如图 17-24 所示的【时间序列建模器】对话框【选项】选项卡，可以设置预测期、指定缺失值的处理、设置置信区间的宽度、为模型的标识指定一个自定义的前缀，以及为自相关设置滞后显示的数量。



图 17-23 【时间序列建模器】对话框【保存】选项卡

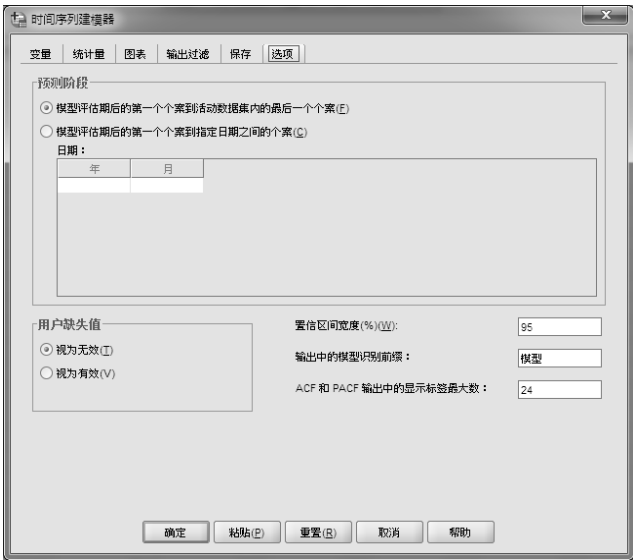


图 17-24 【时间序列建模器】对话框【选项】选项卡

到实际数据集的最后一个观测作为预测期。当估计期先于当前工作数据文件的最后一个样品前结束，又要预测到最后一个样品时选择该项，主要用来为延续期产生预测，允许同期的模型预测值同实际值的子集进行比较。

1. 定义预测期

在【预测阶段】栏中选择预测期的位置。通常预测期从估计期(用来确定模型的样品集)结束后的第一个样品开始一直到当前工作数据文件中的最后一个样品或使用者指定的日期时结束。在默认状态下，估计期结束在当前工作数据文件中的最后一个样品，但它可以在【数据】菜单中的【选择个案】对话框中通过选择【基于时间或个案全距】选项来改变。

(1)【模型评估期后的第一个个案到活动数据集内的最后一个个案】。用估计期结束后的第一个观测

(2) **【模型评估期后的第一个个案到指定日期之间的个案】**。要求指出预测期的结束点,主要用来产生超出实际序列最后范围的预测。在**【日期】**栏中所有单元格输入日期值。如果当前工作的数据文件中没有定义日期说明,那么**【日期】**栏中只显示**【观察】**一列。输入预测期结束点相应样品的行数(同在**【数据编辑器】**中显示的一样)。

在**【日期】**栏若显示**【循环】**则与当前工作的数据文件里循环\_变量的值有关。

## 2. 在**【用户缺失值】**栏中选择对用户缺失值的处理方法

(1) **【视为无效】**。用户缺失值当作系统缺失值处理。

(2) **【视为有效】**。用户缺失值当作有效值处理。

## 3. 定义置信区间

在**【置信区间宽度(%)】**框中,可以指定任意小于 100 的正数。为模型预测计算置信区间和残差自相关。在默认状态下,使用 95%的置信区间。

## 4. 为输出中的模型标识定义前缀

在**【输出中的模型识别前缀】**框中,可以输入前缀或保留模型的默认名。在**【变量】**对话框中指定每个因变量的估计模型。模型用唯一名字区分,名字由定制的前缀和整数后缀组成。

## 5. 定义 ACF 和 PACF 输出中显示的最大滞后数

在**【ACF 和 PACF 输出中显示的最大滞后数】**框中,可以设置在表和自相关与偏自相关图中显示的最大滞后数。

# 17.3.8 时间序列分析实例

**【例 3】** 仍以 17.2.2 节中存放在数据文件 data17-02 中的 1999—2003 年 85 个地区宽带供货商每月的国家宽带服务用户数量的数据文件为例,试用**【专家建模器】**对每个地区宽带供货商每月的国家宽带服务用户数量的数据进行时间序列分析。

**注意:** 变量名须是英文,否则不能作超出数据文件长度的预测。

打开数据文件 data17-02,具体操作步骤如下:

(1) 按**【分析→预测→创建模型】**顺序单击菜单项,打开**【时间序列建模器】**对话框**【变量】**选项卡,见图 17-13。

(2) 在**【变量】**表中选择 Market\_1~Market\_85,即供货商 1 的用户数到供货商 85 的用户数共 85 个变量,并将其移入**【因变量】**框中。要求拟合 85 个模型。

(3) 在**【方法】**下拉列表中选择**【专家建模器】**。

(4) 单击**【条件】**按钮,打开**【时间序列建模器:专家建模器条件】**对话框,见图 17-14。虽然当前周期是 12,但作出序列图后,可以知道 85 个供货商的用户数的时间序列中同样不存在季节性因素的影响,所以可以不考虑季节模型,故在**【模型类型】**栏中,撤销**【专家建模器考虑季节性模型】**选项。这样使用**【专家建模器】**研究时可以减少建模占用的计算机存储空间和计算时间。

(5) 单击**【继续】**按钮,返回**【时间序列建模器】**对话框**【变量】**选项卡。

(6) 单击**【选项】**选项卡,见图 17-24。

在**【预测阶段】**栏中选择**【模型评估期后的第一个个案到指定日期之间的个案】**。在**【日**

期】栏的【年】下面输入“2004”，在【月】下面输入“3”。设置的预测期将从 2004 年 1 月到 2004 年 3 月。其他采用系统默认选项。

(7) 单击【统计量】选项卡，见图 17-20。

选择【显示预测值】，为每个因变量序列产生一张预测值表。

在【比较模型的统计量】栏中，使用【拟合变量】默认选项产生拟合统计量汇总表。

(8) 单击【图表】选项卡，见图 17-21。

由于我们对存储预测值为一个新变量比产生预测图更感兴趣，所以，在【单个模型图】栏中撤销【序列】选项，禁止为每个模型产生序列图。

在【模型比较图】栏中选择【平均绝对误差百分比(MAPE)】和【最大绝对误差百分比(MaxAPE)】。

绝对误差百分比是因变量序列同它的模型预测序列间有多少差异的测度。通过检查所有模型的均数和最大值，可以得到在预测中是否有不确定性的迹象。因此查看百分比误差的概要图，而不是绝对误差图是明智的，因为因变量序列代表大小不同的市场的用户数。

(9) 单击【保存】选项卡，见图 17-23。在【保存变量】栏的【变量】表中选择保存【预测值】，并使用 Predicted 作为变量名前缀，否则软件不认可。

单击【浏览】按钮，设置存储位置，输入 XML 格式文件名“model17-02”。

(10) 单击【确定】按钮，提交系统运行。得到输出结果，见表 17-4～表 17-7 和图 17-21～图 17-23。

(11) 结果解释。

表 17-4 模型描述显示了最佳拟合各供货商用户数的时间序列模型(只选取其中部分供货商)。

第一列显示时间序列变量、供货商用户数的模型编号，第二列显示其对应的最佳拟合模型名称，参见 17.3.2.2、17.3.2.3 节的相关内容。

图 7-25 中给出的直方图显示了所有模型的平均绝对误差百分比频数。由于大部分模型的平均绝对误差百分比在 0.8～1.0 之间，它表明所有模型显示了大概 1% 的平均不确定性。

图 17-26 给出的直方图显示了所有模型的最大绝对误差百分比频数，它对设想预测的最坏情况方案是有用的。它显示每个模型的最大绝对误差百分比落在 1%～5% 的范围中。

这些值能否代表可接受的不确定性的量呢？这要取决于个人的商业直觉，因为可接受的风险将因问题而改变。

表 17-5 列出了模型拟合的各种统计量，从左至右各列依次为拟合统计量、平均数、标准误差、最小值、最大值、百分比(5%、10%、25%、50%、75%、90%、95%)。第一列列出 8 种拟合优度测度，其余各列是这些拟合优度测度统计量的计算结果。

通常重点关注两个统计量：MAPE 平均绝对误差百分比和 MaxAPE 最大绝对误差百分比。例如，模型 95% 的 MaxAPE 的值为 3.456%，所有模型的 MAPE 从最小值 0.647% 到最大值 1.007% 之间变化，所有模型的 MaxAPE 从最小值 1.708% 到最大值 4.765% 之间变化。因此，在各个模型的预报中平均不确定性大约为 1%，最大不确定性在 2.5% 左右(MaxAPE 的平均值)，以及一个大约 4.8% 的最坏情况推测。这些值是否能代表一个可接受的不确定性的量取决于愿意去接受的风险的程度。

表 17-4 模型描述

模型描述			
模型ID		模型类型	
供货商10的用户数	模型_1	ARIMA(1,1,0)	
供货商11的用户数	模型_2	Brown	
供货商12的用户数	模型_3	Brown	
供货商13的用户数	模型_4	ARIMA(1,2,0)	
供货商14的用户数	模型_5	Brown	
供货商15的用户数	模型_6	ARIMA(1,1,0)	
供货商16的用户数	模型_7	ARIMA(0,1,0)	
供货商17的用户数	模型_8	ARIMA(0,1,0)	
供货商18的用户数	模型_9	Brown	
供货商19的用户数	模型_10	Holt	
供货商1的用户数	模型_11	Brown	

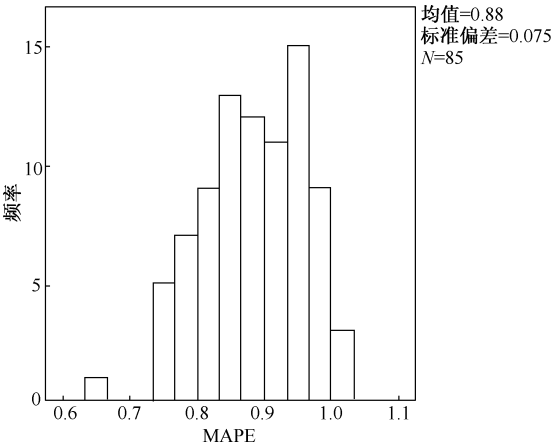


图 17-25 平均绝对误差百分比频数图

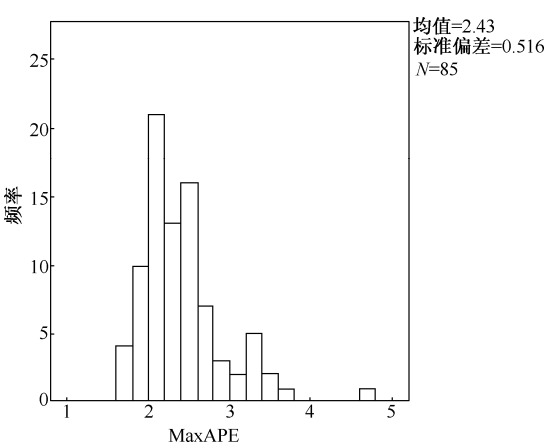


图 17-26 最大绝对误差百分比频数图

表 17-5 模型拟合

拟合统计量	均值	SE	最小值	最大值	百分位						
					5	10	25	50	75	90	95
平稳的 R 方	.183	.144	-2.665E-015	.629	-5.995E-016	.000	.068	.188	.247	.376	.478
R 方	.999	.000	.998	1.000	.998	.999	.999	.999	.999	1.000	1.000
RMSE	177.951	138.821	42.088	737.540	51.471	58.625	90.316	135.016	195.507	402.111	480.796
MAPE	.883	.075	.647	1.007	.748	.775	.831	.885	.946	.980	.991
MaxAPE	2.426	.516	1.708	4.765	1.798	1.937	2.097	2.307	2.604	3.260	3.456
MAE	139.634	106.313	34.033	589.708	40.608	46.836	71.445	107.377	150.785	316.144	343.171
MaxAE	456.170	378.664	108.378	1813.295	117.841	137.366	206.780	340.308	517.760	1070.229	1356.113
正态化的 BIC	10.003	1.316	7.618	13.345	7.994	8.256	9.111	9.902	10.619	12.062	12.445

表 17-6 所示是模型统计数据，从左至右各列依次是模型名称，预测因子的数量，模型拟合统计(平稳 $R^2$ )，Ljung-Box Q(18)统计量、自由度、显著性概率值，异常值数量。第一列依次列出了 85 个供货商用户数的模型，其余各列为对应的计算结果。由于该表很大，限于篇幅，只列出了一部分模型的计算结果。

表 17-7 所示是前两个模型(85 个模型)在指定的预测期 2004 年 1~3 月的预测值。UCL 和 LCL 分别为预测值置信区间(系统默认 95%)上限和下限。

表 17-6 模型统计数据(部分)

模型统计量						
模型	预测变量数	模型拟合统计量	Ljung-Box Q(18)			离群值数
		平稳的 R 方	统计量	DF	Sig.	
供货商10的用户数-模型_1	0	.330	14.768	17	.612	0
供货商11的用户数-模型_2	0	.029	21.373	17	.210	0
供货商12的用户数-模型_3	0	.194	27.548	17	.051	0
供货商13的用户数-模型_4	0	.400	10.053	17	.901	0
供货商14的用户数-模型_5	0	.052	17.099	17	.448	0
供货商15的用户数-模型_6	0	.188	21.825	17	.192	0
供货商16的用户数-模型_7	0	9.992E-016	29.880	18	.039	0
供货商17的用户数-模型_8	0	4.441E-016	16.021	18	.591	0
供货商18的用户数-模型_9	0	.027	27.301	17	.054	0
供货商19的用户数-模型_10	0	.243	24.351	16	.082	0
供货商1的用户数-模型_11	0	.245	10.663	17	.874	0

图 17-27 所示是数据窗中根据预测模型产生的预测值新变量。每个新变量包含估计期的模型预测值(从 1999 年 1 月到 2003 年 12 月),还包括指定的预测期 2004 年 1~3 月的预测值,因此增加了 3 个新样品。可以根据估计期的预测值观察模型拟合的优劣。

每个供应商的模型都有 1 个新变量,共 85 个,图中只显示了 11 个。

表 17-7 预测部分结果

预测				
模型		一月 2004	二月 2004	三月 2004
供货商10的用户数-模型_1	预测	23260	23508	23768
	UCL	23565	24076	24577
	LCL	22954	22941	22960
供货商11的用户数-模型_2	预测	14202	14344	14486
	UCL	14461	14879	15352
	LCL	13944	13809	13620

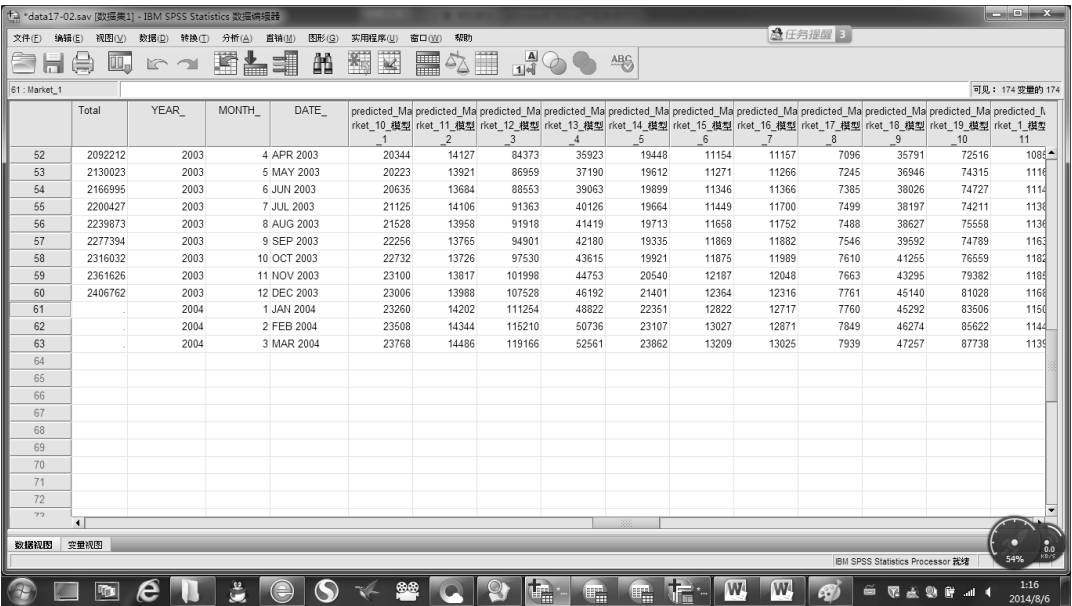


图 17-27 数据窗中根据预测模型产生的预测值的新变量

17.4 应用时间序列模型

【应用模型】过程是与【创建模型】过程相关联的一个过程,它是【创建模型】过程的延续,它将【创建模型】过程中建立的模型,应用于原建模的时间序列得到延长或修正后的更新预测中。

像某公司各月的销售量这样的时间序列数据,时间序列的长度会随着销售时间的变化而变化,存放这种时间序列的数据文件,也会随着时间变化而不断增加新增的序列数据。如果对这样数据文件中的时间序列已在某个时期使用过专家建模器创建了模型,并将建模结果保存到后缀为.XML 的外部文件,则在原建模的数据文件中新增时间序列的数据,或对其中时间序列的某些数据进行修正后,想要得到更新的时间序列的预测值时,就无须再用上述的【创建模型】过程重新建模,而可以直接使用【应用模型】过程,再调用在【创建模型】过程中已保存的建模文件,即可得到时间序列更新的预测值。

需要注意的是,【应用模型】过程中使用的数据文件中的变量名、变量标签名等应与【创建模型】过程中使用的数据文件中的变量名、变量标签名等应完全一致,否则可能无法得到超出数据文件中最后一个个案的指定的延长期的预测结果。

## 17.4.1 应用时间序列模型过程

按【分析→预测→应用模型】顺序单击菜单项,打开如图 17-28 所示的【应用时间序列模型】对话框。



图 17-28 【应用时间序列模型】对话框

量或自变量而言,预测都不考虑历史数据。如果想使用历史数据去影响预测,应选择【从数据中重新评估】。另外,预测不考虑预测期中因变量序列的值,但考虑预测期中自变量的值。如果有更多的因变量序列的现值,并想要在预测中包括它们,那么必须选择【从数据中重新评估】,调整估计期来包含这些值。

②【从数据中重新评估】。根据当前工作的数据集中的数据重新估计模型参数。模型参数的重新估计不改变模型结构。例如,ARIMA(1,0,1)模型依旧,但要重新估计自回归和移动平均参数。重新估计时不进行异常值的重新检测。如果有异常值话,那是来自模型文件的。

【估计阶段】栏显示的是默认的估计期。它定义了使用于重新估计模型参数的样品集。默认估计期包括当前工作数据集中的所有样品。可以使用【数据】菜单【选择个案】功能中的【基于时间或个案全距】重新定义估计期。估计期取决于有效数据,程序使用的估计期可以因模型而异,因此会与显示值不同。对一个给定的模型,真实的估计期是从模型的因变量中消除在指定估计期的开始和结束中发生的任何相邻的缺少值之后剩下的时段。

(3) 在【预测阶段】栏中定义预测期。具体内容参见 17.3.7 节中的相关内容。

(4)【统计量】、【图表】、【输出过滤】、【保存】、【选项】选项卡中的内容参见 17.3.3~17.3.7 节中的相关内容。

## 17.4.2 应用时间序列模型分析实例

【例 4】以 17.3 节中建立的 model17-02.xml 为模型基础,使用在数据文件 data17-02 基础上补充进 2004 年 1~3 月各供货商的实际用户数据后形成的新数据文件 data17-03 进行预测 2004 年 4~6 月的各月用户数,以此来说明【应用模型】的使用方法。

在 SPSS 中,使用【应用模型】的基本步骤如下:

(1) 在 SPSS 数据编辑窗中,打开数据文件 data17-03.sav。

(2) 按【分析→预测→应用模型】顺序单击菜单项, 打开【应用时间序列模型】对话框, 见图 17-28。

单击【浏览】按钮, 然后选择数据盘中的“model17-02.xml”。

为使时间序列的新值加入到预报中, 【应用模型】程序必须重新估计模型参数, 故选择【从数据中重新评估】。由于模型的结构仍然是一样的, 因此计算重新估计的时间远远少于原先建模的计算时间。

用来重新估计的样品集需要包括新数据。如果使用默认的【第一个个案到最后一个个案】估计期, 将是保险的。如果有时需要对除了系统默认外的事情设置估计期, 则可以通过在【数据】菜单下的【选择个案】对话框中选择【基于时间或个案全距】来完成。

在【预测阶段】栏中选择【模型评估期后的第一个个案到指定日期之间的个案】。

在【日期】栏的【年】框中输入“2004”, 在【月】框中输入“6”。

数据集包含 1999 年 1 月至 2004 年 3 月的数据。用当前的设置, 预测期将是 2004 年 4 月到 2004 年 6 月。

(3) 单击【保存】选项卡, 在【保存】列中选择【预测值】, 在【变量名的前缀】列中输入“Predicted”。

模型预测值将被作为新变量存储在当前工作的数据集中, 新变量使用前缀 Predicted。

(4) 单击【图表】选项卡, 由于对存储预测值作为新变量比产生预测图更感兴趣, 所以在【单个模型图】项中撤销【序列】选项。这可以阻止为每个模型产生序列图。

(5) 单击【确定】按钮运行, 在输出窗中得到与表 17-4~表 17-6 类似的结果, 以及在当前工作文件中得到与图 17-27 类似的结果。有关它们的解释, 参见 17.3.8 节中的相关图、表的解释。

## 17.5 自 相 关

自相关系数值度量了不同时间点上的观察值之间的相关程度。它常用来洞悉产生数据的概率模型。解释自相关系数值集合的一个有效工具是自相关图。自相关图包括自相关函数(ACF)图和偏自相关函数(PACF)图两种。

使用【自相关】程序可以绘制 ACF 图及一个或一个以上序列的 PACF 图。需要注意, 【自相关】过程只适合于时间序列数据。

### 17.5.1 自相关系数与偏自相关系数的计算

设第  $i$  个输入序列的观察值为  $x_i (i=1, \dots, n)$ , 第  $k$  个滞后样本自相关系数为  $r_k$ , 第  $k$  个滞后样本偏自相关系数为  $\hat{\phi}_{kk}$ 。

如果没有遇到缺失值, 则使用没有缺失值情形中提供的公式计算; 如果出现缺失值, 则使用含有缺失值情形中提供的修正公式计算。

#### 1. 没有缺失值情形

(1) 样本自相关系数的计算公式为

$$r_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

式中,  $\bar{x}$  是  $n$  个观察值的均值。

## (2) 样本自相关系数标准误的计算公式

根据对自相关的不同假定, 有两个计算  $r_k$  的标准误的公式。在 MA 过程的阶数是  $k-1$  为真的假定下,  $r_k$  的近似方差为

$$\text{VAR}(r_k) \cong \ln \left( 1 + 2 \sum_{l=1}^{k-1} r_{2l} \right)$$

在该过程为白噪声的假定下,  $r_k$  的近似方差为

$$\text{VAR}(r_k) \cong \ln \left( \frac{1}{n} \left( \frac{n-k}{n+2} \right) \right)$$

式中,  $r_k$  的标准误是以上方差的平方根。

## (3) Box-Ljung 统计量

在滞后  $k$  处, Box-Ljung 统计量被定义为

$$Q_k = n(n+2) \sum_{l=1}^k \frac{r_l^2}{n-l}$$

当  $n$  大时,  $Q_k$  服从自由度为  $k-p-q$  的卡方分布,  $p$  和  $q$  分别为自回归和移动平均的阶数。用自由度为  $k-p-q$  的卡方分布来计算  $Q_k$  的显著性水平。

## (4) 样本偏自相关系数为

$$\begin{aligned} \hat{\phi}_{11} &= r_1 \\ \hat{\phi}_{22} &= (r_2 - r_1^2) / (1 - r_1^2) \\ \hat{\phi}_{kj} &= \hat{\phi}_{k-1,j} - \hat{\phi}_{kk} \hat{\phi}_{k-1,k-j} \quad (k=2, \dots; j=1, 2, \dots, k-1) \\ \hat{\phi}_{kk} &= \left( r_k - \sum_{j=1}^{k-1} \phi_{k-1,j} r_{k-j} \right) / \left( 1 - \sum_{j=1}^{k-1} \phi_{k-1,j} r_{k-j} \right) \quad (k=3, \dots) \end{aligned}$$

## (5) 样本偏自相关系数的标准误。

在 AR(P) 模型是相关及  $p \leq k-1$  的假定下, 样本偏自相关系数为

$$\hat{\phi}_{kk} \cong N(0, 1/n)$$

因此, 偏自相关系数的方差为  $\text{VAR}(\hat{\phi}_{kk}) \cong \frac{1}{n}$ 。其标准误是该方差的平方根。

## 2. 时间序列含有缺失值的情形

如果在  $x$  中有缺失值, 则以下计算的统计量是不同的。首先, 定义

$$\begin{aligned} \bar{x} &= \text{非缺失值 } x_1, \dots, x_n \text{ 的平均值} \\ a_i &= \begin{cases} x_i - \bar{x} & \text{如果 } x_i \text{ 不是缺失值} \\ \text{系统缺失值} & \text{如果 } x_i \text{ 是缺失值} \end{cases} \end{aligned}$$

对于  $k=0, 1, 2, \dots$  及  $j=1, \dots, n$  有

$$b_j^{(k)} = \begin{cases} a_j a_{j+k} & \text{如果都不是缺失值} \\ \text{系统缺失值} & \text{其他} \end{cases}$$

式中,  $m_k =$  在  $b_1^{(k)}, \dots, b_{n-k}^{(k)}$  中非缺失值的数量;  $m_0 =$  在  $x$  中非缺失值的数量。

## (1) 样本自相关系数的修正计算公式为

$$r_k = \frac{\text{非缺失值 } b_1^{(k)}, \dots, b_{n-k}^{(k)} \text{ 之和}}{\text{非缺失值 } b_1^{(0)}, \dots, b_{n-k}^{(0)} \text{ 之和}}$$



(2) 样本自相关系数的标准误为

$$se(r_k) = \sqrt{\frac{1}{m_0} \left( 1 + \sum_{l=1}^{k-1} r_l^2 \right)} \quad (\text{MA 假定})$$

$$se(r_k) = \sqrt{\frac{m_k}{(m_0 + 2)m_0}} \quad (\text{白噪声})$$

(3) Box-Ljung 统计量修正计算公式为

$$Q = m_0(m_0 + 2) \sum_{l=1}^k \frac{r_l^2}{m_l}$$

(4) 样本偏自相关系数的标准误为

$$se(\hat{\phi}_{kk}) = \sqrt{\frac{1}{m_0}}$$

## 17.5.2 自相关图

要解释自相关图的含义是一件很困难的事情，这里只给出一般的基本概念。

### 1. 随机序列

如果时间序列是完全随机的，则当时间序列的长度  $N$  很大时，此时得到的自相关系数值近似服从均数为 0，方差为  $1/N$  的标准正态分布。根据置信区间的理论可知，在自相关图上，95% 的自相关系数值应出现在  $\pm 1.96N^{-1/2}$  之间，这就是说，每 20 个自相关系数值至少应有 19 个自相关系数值位于这个区间内，只有 1 个可能例外。但当随机序列中有异常值存在时，很可能看到不止一个自相关系数值出现在该区间之外。当时间序列中存在趋势或季节效应时，也能看到这种情况。这就要求在作自相关分析之前，要用以上所讲到的内容首先处理时间序列中的异常值，并从专业的角度来预判时间序列中是否存在趋势和季节效应，否则将给后续分析带来巨大的麻烦。

### 2. 短期相关

平稳序列常显示出短期相关，其明显的特征是，第一个自相关系数值很大，其后的自相关系数值大于 0 但逐渐减小，较长滞后的相关系数近似趋向于 0。对这种类型的时间序列可用自回归模型来加以拟合。

### 3. 非平稳序列

如果时间序列含有趋势，除非滞后值很大，否则自相关系数值是不会下降为 0 的。这是由于趋势的存在使得总均值一侧的观察值有大量后续的观察值也倾向于在均值的同一侧。对此类型的相关图，因为趋势支配所有其他特征，所以很难推出什么结论。因此在计算自相关系数值之前，要用前面提到的差分等手段消除时间序列中的趋势。

### 4. 季节波动

含有季节波动的时间序列在自相关图上表现为出现相同频率的振荡。对于逐月观测的时间序列来说，第 6 个自相关系数值将是绝对值大而本身是负的，第 12 个自相关系数值将是大而正的。当自相关系数值形成正弦模型时，也会出现同样的规律。含有季节波动的数据，在数据

的时序图上有十分清晰的表现。但对这种类型的季节数据,相关图没有提供更多的额外信息。如果将季节变化从数据中去除,如用各点的时间序列数据减去各个季节对应点上的平均值后得到的消除季节影响后序列的自相关图,就可以进行分析了。

## 5. 交错序列

如果交错趋势存在时间序列中,则相关图也会出现交错的趋势。例如,第一个自相关系数值为负,则第二个自相关系数值为正。这是由于相邻的观察值总是总均值的两侧所造成的,这样凡滞后 2 的观察值总在总均值的同一侧,故第二个自相关系数值总是大于 0 的。

解释自相关图需要大量的实际工作经验,多学、多看、多做是掌握自相关图解释的唯一途径。至于各种时间序列的具体的特征图,不是本书的重点,在一般的时间序列分析书籍中都有介绍,在此不一一列出。

## 17.5.1 自相关分析过程

按【分析→预测→自相关】顺序单击菜单项,打开如图 17-29 所示的【自相关】对话框。

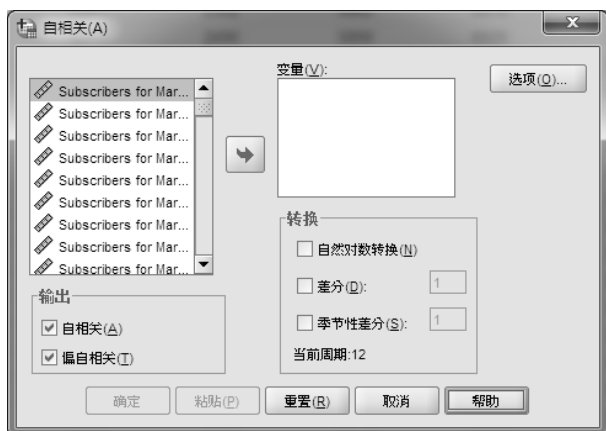


图 17-29 【自相关】对话框

### 1. 定义变量

在源变量表中选择一个或多个数值型变量,送入【变量】框中。

### 2. 在【输出】栏中定义显示函数

①【自相关】。序列同滞后 1 或多个样品值的相关值。选择该项,计算 1, 2, ..., 直到一个指定数的滞后的自相关。

②【偏自相关】。计算在干涉滞后相关的影响被消除之后,序列同滞后 1 个或多个样品值的相关值。

显示偏自相关需要解方程组,方程组的规模随滞后数的增大而增大。对高阶滞后(大于 24)要求作偏自相关要注意,即使在高速计算机上,也会比求自相关要花更长的时间。假如有季节因素影响的序列需要看看高阶滞后,则在确信序列是平稳序列之前建议先看看自相关,然后再要求作偏自相关。

### 3. 在【转换】栏中选择数据转换方法

时间序列数据转换的目的是要使偏态分布的序列变成对称分布的序列,消除序列中的异方差性,使变量间的非线性变成线性,使非平稳的序列变成平稳序列等,以便满足所选模型的要求,有利于进一步的分析。

除系统默认不作转换外,另外还有以下 3 个可选项:

①【自然对数转换】。对方差非平稳的序列,如对时间序列的散点图有指数曲线趋势的时间序列,可选择该项进行转换。

②【差分】转换。差分是通过用时间序列的逐项相减来消除前后期数据的相关性的方法,可剔除序列中的趋势性。对含有一般趋势的时间序列,可选择该项进行转换。需要在右边的文

本框中输入正整数值，作为指定的差分阶数。一般而言，具有线性趋势的时间序列，只需进行 1 阶差分即可，而具有  $d$  阶多项式趋势的时间序列，需进行  $d$  阶差分。

③【季节性差分】转换。在含有周期性趋势的时间序列中，把每个观察值与下一个周期相应时刻的观察值相减，所得差值称为季节性差分。对含有周期性趋势的时间序列，可选择该项进行转换。需要在右边的文本框中输入正整数值，作为指定的季节性差分阶数。

在【当前周期】后面显示当前周期，为正整数值，它是由工作数据集中定义的日期变量的周期决定的。

#### 4. 单击【选项】按钮

打开如图 17-30 所示的【自相关：选项】对话框。

(1)【最大延迟数】框。默认值为 16。可以重新输入最大滞后数。

(2)在【标准误差】栏中选择计算标准误差的方法。如果在主对话框中，撤销【自相关】选项，则该栏无效。

①【独立模型】。假设潜在过程是白噪声时的标准误差。

②【Bartlett 的近似值】。用近似值计算标准误差，适用于序列描述  $k-1$  阶移动平均过程。用这种方法，标准误差随滞后的增加而变大。

(3)【在周期延迟处显示自相关】。显示周期性滞后处的自相关，如果已经定义了季节性，可以选择该项。

单击【继续】按钮，返回主对话框。在主对话框中单击【确定】按钮，运行过程。



图 17-30 【自相关：选项】对话框

### 17.5.3 自相关分析实例

【例 5】数据文件 data17-04 中的变量 sales 为某公司 1986—1997 年间各季度某商品的销售量数据，用自相关法对其进行统计学分析。打开数据文件 data17-04，分析步骤如下：

- (1) 按【分析→预测→自相关】顺序单击菜单项，打开【自相关】对话框。
- (2) 在源变量表中选择销售量 sales 作为分析变量，送入【变量】框中。
- (3) 其他保持系统默认选项，单击【确定】按钮，运行过程，输出窗中出现表 17-8～表 17-11 及图 17-31、图 17-32 所示的结果。
- (4) 结果解释。

表 17-8 列出了模型的基本描述，从上至下依次为模型名称(MOD\_1)、序列名 1(sales)、转换(无)、非季节性差分(0)、季节性差分(0)、季节期间的长度(4)、最大滞后数(16) (此为系统默认值)、为计算自相关的标准误而假定的过程(独立性(白噪音))、显示并绘图(所有滞后)。

表 17-9 所示为样品处理摘要。

表 17-10 所示为是自相关计算结果，从左至右依次列出的是：滞后数、自相关系数值、标准误差、Box-ljung 统计量(值、自由度、原假设成立的概率值)。在原假设(假设基本过程是独立的，也即假定时间序列所反映的随机过程是白噪声)成立的前提下，出现大于等于目前统计量值的概率值都小于 0.05，所以全部自相关均有显著性意义。

表 17-11 从左至右依次列出的是：滞后数、偏自相关、标准误差的计算结果。

图 17-27 所示是对应于表 17-10 自相关系数值的自相关图。图 17-28 所示是对应于表 17-11 偏自相关系数值的偏自相关图。

表 17-8 模型描述

模型描述		
模型名称	MOD_1	
序列名	1	sales
转换	无	
非季节性差分		0
季节性差分		0
季节性期间的长度		4
最大滞后数		16
为计算自相关的标准误而假定的过程	独立性（白噪音） <sup>a</sup>	
显示并绘图	所有滞后	

正在应用来自 MOD\_1 的模型指定。  
a. 不适用于计算偏自相关的标准误。

表 17-9 样品处理摘要

个案处理摘要		sales
序列长度		48
缺失值数	用户缺失	0
	系统缺失	0
有效值数		48
可计算的第一滞后数		47

表 17-10 自相关计算结果

自相关图					
序列: sales					
滞后	自相关	标准误差 <sup>a</sup>	Box-Ljung 统计量		
			值	df	Sig. <sup>b</sup>
1	.432	.140	9.539	1	.002
2	.188	.138	11.377	2	.003
3	.317	.137	16.729	3	.001
4	.799	.135	51.565	4	.000
5	.288	.134	56.207	5	.000
6	.068	.132	56.471	6	.000
7	.164	.131	58.038	7	.000
8	.598	.129	79.517	8	.000
9	.143	.127	80.778	9	.000
10	-.049	.126	80.927	10	.000
11	.047	.124	81.070	11	.000
12	.451	.122	94.650	12	.000
13	.061	.121	94.902	13	.000
14	-.111	.119	95.773	14	.000
15	-.041	.117	95.896	15	.000
16	.318	.115	103.459	16	.000

a. 假定的基础过程是独立性（白噪音）。  
b. 基于渐近卡方近似。

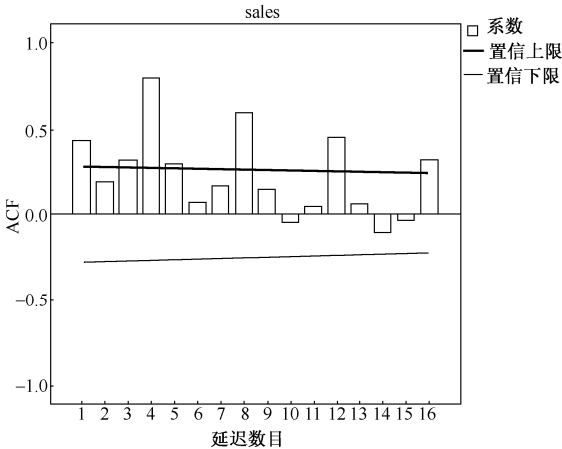


图 17-31 自相关图

表 17-11 偏自相关表

偏自相关		
序列: sales		
滞后	偏自相关	标准误差
1	.432	.144
2	.001	.144
3	.289	.144
4	.752	.144
5	-.592	.144
6	.127	.144
7	-.051	.144
8	-.034	.144
9	-.064	.144
10	-.010	.144
11	.069	.144
12	-.039	.144
13	.057	.144
14	-.054	.144
15	-.076	.144
16	-.035	.144

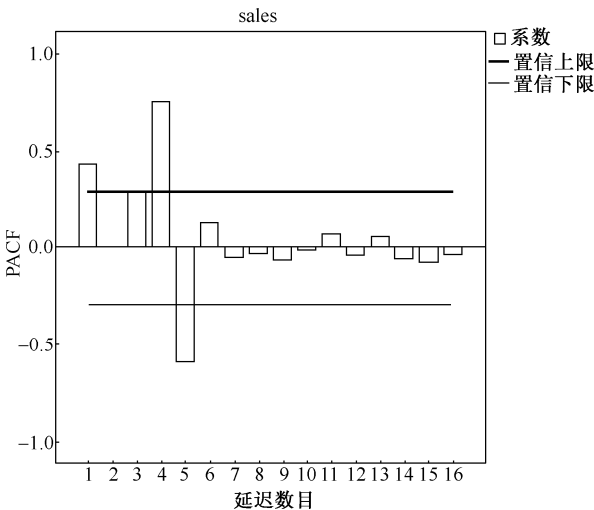


图 17-32 偏自相关图

在滞后 4 处的重要的顶点暗示在数据中存在周期为 4(4 个季度)的季节成分。检查偏自相关函数图同样可得到这个十分明确的结论。

## 17.6 季节分解法

在实际工作中,经常会遇到按日、周、月、季或年记录的数据资料,如每天新生儿出生的情况、某产品每月的销售量、每年 GDP 的增长率等。这些资料通过自相关分析可能符合季节性分布,对这些有随机变异、长期趋势、季节效应或周期变动的时间序列资料,可以使用季节分解法对其进行分析,从而得到有意义的结果。

### 17.6.1 季节分解法模型

#### 1. 模型种类

模型分为两类:乘法模型和加法模型。

(1) 乘法模型:  $X_t = TC_t S_t I_t$  ( $t=1, \dots, n$ )。

(2) 加法模型:  $X_t = TC_t + S_t + I_t$  ( $t=1, \dots, n$ )。

在上述公式中,  $TC_t$  是“趋势循环”成分;  $S_t$  是“季节”成分;  $I_t$  是“无规律”或“随机”成分;  $X_t$  为时间序列;  $t$  为时间点;  $n$  为时间序列的长度,下同。

估计季节成分的程序是:

(1) 使用移动平均法平滑时间序列:移动平均序列反映趋势循环成分。

(2) 如果模型是乘性的,则使用通过平滑值划分的初始序列获取季节不规则成分;如果模型是加性的,则通过从原始序列中减去平滑值来获取季节不规则成分。

(3) 如果模型是乘性的(加性的),则通过为周期的每个单元计算指定季节有关的调和均数,从季节不规则成分中分离季节成分。

#### 2. 移动平均时间序列

基于指定的方法和周期  $p$ ,  $X_t$  的移动平均序列  $Z_t$  按如下定义:

①  $P$  为偶数,所有点权重相等,即

$$Z_t = \begin{cases} \sum_{j=t-\frac{p}{2}}^{t+\frac{p}{2}-1} X_j / p & t = \frac{p}{2} + 1, \dots, n - \frac{p}{2} + 1 \\ \text{系统缺失值} & \text{其他} \end{cases}$$

②  $P$  为偶数,所有点权重不等,即

$$Z_t = \begin{cases} \left( X_{t-\frac{p}{2}} + X_{t+\frac{p}{2}} \right) / 2p + \sum_{j=t-\frac{p}{2}+1}^{t+\frac{p}{2}-1} X_j / p & t = \frac{p}{2} + 1, \dots, n - \frac{p}{2} + 1 \\ \text{系统缺失值} & \text{其他} \end{cases}$$

③  $P$  为奇数,即

$$Z_t = \begin{cases} \sum_{j=t-\lfloor \frac{p}{2} \rfloor}^{t+\lfloor \frac{p}{2} \rfloor} X_j / p & t = \lfloor \frac{p}{2} \rfloor + 1, \dots, n - \lfloor \frac{p}{2} \rfloor \\ \text{系统缺失值} & \text{其他} \end{cases}$$

3. 比率或差分(季节不规则成分)

- (1) 乘法模型:  $SI_t = \begin{cases} \text{系统缺失值} & \text{如果 } Z_t = \text{系统缺失值} \\ (X_t / Z_t) \times 100 & \text{其他} \end{cases}$
- (2) 加法模型:  $SI_t = \begin{cases} \text{系统缺失值} & \text{如果 } Z_t = \text{系统缺失值} \\ X_t + Z_t & \text{其他} \end{cases}$

4. 季节因素(季节成分)

(1) 乘法模型为

$$F_t = \begin{cases} \text{调和均数}(SI_{t+p}, SI_{t+2p}, \dots, SI_{t+qp}) & 1 \leq t \leq L - \lfloor \frac{L}{P} \rfloor P \\ \text{调和均数}(SI_{t+p}, SI_{t+2p}, \dots, SI_{t+(q-1)p}) & L - \lfloor \frac{L}{P} \rfloor P < t \leq \lfloor \frac{p}{2} \rfloor \\ \text{调和均数}(SI_t, SI_{t+p}, \dots, SI_{t+(q-1)p}) & \lfloor \frac{p}{2} \rfloor < t \leq p \end{cases}$$

式中,  $L = n - \frac{p}{2} + 1$ ,  $q = \lfloor \frac{L}{p} \rfloor$ , 如果  $P$  为偶数, 所有点权重相等;  $L = n - \lfloor \frac{p}{2} \rfloor$ ,  $q = \lfloor (n - p / 2) / p \rfloor$ , 其他情况。

并且在排除最小值和最大值之后, 序列的调和均数等于序列的平均值。季节因素定义如下:

$$SAF_t = F_t \frac{100p}{\sum_{t=1}^p F_t} \quad (t = 1, \dots, p)$$

(2) 加法模型。 $F_t$  被定义为上面显示的序列的算术平均数, 则

$$SAF_t = F_t - \bar{F}$$

式中,  $\bar{F} = \sum_{t=1}^p F_t / p$ 。

5. 季节性调整序列(SAS)

$$SAS_t = \begin{cases} (X_t / SAF_m) 100 & \text{如果模型是乘性的} \\ X_t - SAF_m & \text{如果模型是加性的} \end{cases}$$

式中,  $m = t - \lfloor t / p \rfloor p$ 。

6. 平滑趋势循环序列

通过对季节调整序列(SAS)使用  $3 \times 3$  移动均数获取平滑趋势循环序列(STC)。因而有

$$STC_t = \frac{1}{9}[(SAS)_{t-2} + 2(SAS)_{t-1} + 3(SAS)_t + 2(SAS)_{t+1} + (SAS)_{t+2}] \quad (t = 2, \dots, n-2),$$

并且在序列的起、止位置上的两个末端点

$$(STC)_2 = \frac{1}{3}[(STC)_1 + (STC)_2 + (STC)_3]$$

$$(STC)_{n-1} = \frac{1}{3}[(STC)_{n-2} + (STC)_{n-1} + (STC)_n]$$

$$(STC)_1 = (STC)_2 + \frac{1}{2}[(STC)_2 - (STC)_3]$$

$$(STC)_n = (STC)_{n-1} + \frac{1}{2}[(STC)_{n-1} - (STC)_{n-2}]$$

## 7. 不规则成分

对于  $t=1, \dots, n$  有

$$I_t = \begin{cases} (SAS)_t / (STC)_t & \text{如果模型是乘性的} \\ (SAS)_t - (STC)_t & \text{如果模型是加性的} \end{cases}$$

## 17.6.2 季节分解法分析过程

(1) 进行季节分解的数据，要有至少包括 4 个完整季节数据的变量。打开【数据】菜单中的【定义日期】对话框，定义时间序列的周期，然后才能进行季节分解。

(2) 按【分析→预测→季节性分解】顺序单击菜单项，打开【周期性分解】对话框，见图 17-33，用来估计时间序列的乘性或加性季节因素。

(3) 指定需要季节分解处理的变量。从源变量框中选择分析的变量，移到【变量】框中。该变量必须包括 4 个完整的季节数据。

(4) 在【模型类型】栏中，根据时间序列构成的特点，选择【乘法】模型或【加法】模型。

(5) 在【移动平均权重】栏中，指定在计算移动平均时如何对待时间序列。

①【所有点相等】。计算周期跨度相等和所有点权重相等时的移动平均，常用于周期是奇数的情形。

②【结束点按 0.5 加权】。用相同跨度(周期+1)和端点权重乘 0.5 计算移动平均。该选项仅当时间序列的周期是偶数时有效。

(6) 【显示对象删除列表】(应为：显示个案明细表)。在输出窗中，输出个案在各种方法分解时的明细表，并可在运算过程中对每个变量生成一行 4 个新序列值作为新变量，存放在当前工作的数据集中。

(7) 单击【保存】按钮，打开【周期：保存】对话框，见图 17-34。



图 17-33 【周期性分解】对话框



图 17-34 【周期：保存】对话框

- ①【添加至文件】。季节分解产生的新序列被作为新变量保存在数据窗中。变量名由 3 部分组成：3 个字母的前缀、下画线、数字。这是系统默认的命名方法。
- ②【替换现有】。季节分解产生的新序列作为临时变量保存在数据窗中，已经存在的临时变量被剔除。变量名由 3 部分构成：3 个字母的前缀、井号和一位数字。
- ③【不要创建】，新序列不添加到数据文件中。
- (8) 单击【确定】按钮，系统立即执行命令。

17.6.3 季节分解法分析实例

【例 6】数据文件 data17-04 中的变量 sales 为某公司 1986—1997 年间各季度某商品的销售额数据，用季节分解法对其进行统计学分析。

- 1) 操作方法
- (1) 按【分析→预测→季节性分解】顺序单击菜单项，打开如图 17-33 所示的对话框。
- (2) 选择销售量 sales 变量进入【变量】框。
- (3) 在【模型类型】栏中选择【乘法】。
- (4) 在【移动平均权重】栏中选择【所有点相等】。
- (5) 使用【周期：保存】对话框中的默认设置。
- (6) 单击【确定】按钮，执行运算。
- 2) 输出结果(见表 17-12、表 17-13)
- 表 17-12 给出了模型描述。
- 表 17-13 列出了季节及其对应的季节因素指数。
- 在数据窗中生成来自给定模型的销售量的误差项(ERR\_1)、季节校准序列(SAS\_1)、季节因素指数(SAF\_1)、季节趋势周期(STC\_1)4 列新数据，见图 17-35。

表 17-12 模型描述

模型描述	
模型名称	MOD_1
模型类型	可乘
序列名称	1 sales
季节性期间的长度	4
移动平均数的计算方法	跨度等于周期，并且所有点具有相同的权重

正在应用来自 MOD\_1 的模型指定。

表 17-13 季节因素

季节性因素	
序列名称: sales	
期间	季节性因素 (%)
1	111.8
2	109.2
3	75.8
4	103.2

	sales	year	quarter	date	ERR_1	SAS_1	SAF_1	STC_1
1	3017.60	1986	1 Q1 1986	0.98512	2698.66383	1.11818	2739.42113	
2	3043.54	1986	2 Q2 1986	1.01355	2787.17429	1.09198	2749.92652	
3	2084.35	1986	3 Q3 1986	0.99748	2763.94146	0.75774	2770.93731	
4	2809.84	1986	4 Q4 1986	0.97090	2722.45957	1.03210	2804.06932	
5	3274.80	1987	1 Q1 1987	1.03047	2928.67985	1.11818	2842.09329	
6	3163.28	1987	2 Q2 1987	1.01125	2896.82826	1.09198	2864.59798	
7	2114.31	1987	3 Q3 1987	0.96847	2790.28293	0.75774	2881.11122	

图 17-35 数据文件中增加的 4 个新变量

17.7 频 谱 分 析

17.7.1 频谱分析概述

1. 频谱分析的作用
- 当把时间序列看作是由不同频率的正弦、余弦波组成时，就可用 Schuster 在 1898 年引入的周期图来进行时间序列分析了。周期图最初是用来检测和估计混在噪声中、频率为已知的正弦分量的振幅。用它提供的方法也可检验序列的随机性。在周期图的基础上，用功率谱可以建立样本谱。同样，样本谱也可用来检验和估计隐含于噪声中未知频率的正弦分



量的振幅。尤其当事先已知频率  $f$  并不具有与序列长度的谐振关系时, 样本谱更是实现上述目的的有力工具。可以证明, 样本谱是自协方差函数估计值的傅里叶余弦变换。这为谱分析理论奠定了基础。

当平稳时间序列的频率、振幅和相位都是随机变化时, 样本谱失去了其应有的作用, 此时用频率强度的均值建立起来的谱分析是最重要的分析工具。

频谱分析过程可用来识别时间序列中的周期行为, 而不是分析一个时间点向下一个时间点的变化, 通常是把分析序列的变化转化成不同频率的周期成分。平稳序列在低频时有更强的周期成分, 随机变化的白噪声遍及所有频率。

SPSS 中的频谱分析可为一个或多个时间序列绘制周期图及谱密度函数估计。

分析变量应该是数字型、平稳的不包含缺失值的时间序列; 应从时间序列中减去任何的非零均值; 应在预测分析前处理缺失值, 方法参见替换缺失值方面的内容。将不稳定序列变成平稳序列的常用方法是差分转换, 请参阅建立时间序列方面的内容。

## 2. 分析中使用的计算公式

### (1) 单变量序列。

对于所有的  $t$ , 序列  $X_t$  的谱密度函数可以表示为

$$X_t = a_0^x + \sum_{K=1}^q [a_K^x \cos 2\pi f_K(t-1) + b_K^x \sin 2\pi f_K(t-1)]$$

$$\text{式中, } t=1, 2, \dots, N; \quad a_0^x = \bar{x}; \quad \bar{x} = \sum_{t=1}^N x_t / N; \quad a_K^x = \frac{2}{N} \left[ \sum_{t=1}^N (X_t \cos 2\pi f_K(t-1)) \right]; \quad b_K^x = \frac{2}{N} \sum_{t=1}^N [X_t \sin 2\pi f_K(t-1)]; \quad f_K = \frac{K}{N}; \quad q = \begin{cases} N/2 & \text{如果 } N \text{ 是偶数} \\ (N-1)/2 & \text{如果 } N \text{ 是奇数} \end{cases}.$$

计算下述统计量。

- ① 频率:  $f_K = K / N (K=1, \dots, q)$ 。
- ② 周期:  $1 / f_K = N / K (K=1, \dots, q)$ 。
- ③ 傅里叶余弦系数:  $a_K^x, K=1, \dots, q$ 。
- ④ 傅里叶正弦系数:  $b_K^x = (a_K^x - ib_K^x)(a_K^x + ib_K^x), K=1, \dots, q$ 。
- ⑤ 周期图谱:  $I_K^x = [(a_K^x)^2 + (b_K^x)^2]N/2, K=1, \dots, p$

谱密度估计为

$$s_K^x = \sum_{j=-p}^p w_j I_{K+j}^x$$

式中,  $2p+1=m$  (跨度的数量);  $I_{-K}^x = I_K^x, K=1, \dots, q, I_0^x = I_1^x, I_K^x = I_{N+1-K}^x (K > q)$ 。  
 $w_{-p}, w_{-p+1}, w_0, w_1, \dots, w_p$  为由不同数据窗口定义的周期图谱权重。

### (2) 双变量序列。

双变量序列  $X_t$  和  $Y_t$  的谱密度函数可以表示为

$$X_t = a_0^x + \sum_{K=1}^q (a_K^x \cos 2\pi f_K t + b_K^x \sin 2\pi f_K t) \quad (t=1, 2, \dots, N)$$

$$Y_t = a_0^y + \sum_{K=1}^q (a_K^y \cos 2\pi f_K t + b_K^y \sin 2\pi f_K t) \quad (t=1, 2, \dots, N)$$

①  $X$  和  $Y$  的互周期图为

$$I_K^{xy} = \frac{N}{2}(a_K^x - ib_K^x)(a_K^y + ib_K^y) = \frac{N}{2}[(a_K^x a_K^y + b_K^x b_K^y) + i(a_K^x b_K^y - b_K^x a_K^y)]$$

② 实部与虚部。

$$\text{实部: } (\text{RC})_K = \frac{N}{2}(a_K^x a_K^y + b_K^x b_K^y)。$$

$$\text{虚部: } (\text{IC})_K = \frac{N}{2}(a_K^x a_K^y - b_K^x b_K^y)。$$

③ 余谱密度估计为

$$C_K = \sum_{j=-p}^p w_j (\text{RC})_{K+j}$$

④ 正交谱估计为

$$Q_K = \sum_{j=-p}^p w_j (\text{IC})_{K+j}$$

⑤ 交叉振幅值为

$$A_K = (Q_K^2 + C_K^2)^{1/2}$$

⑥ 平方一致性值为

$$K_K = \frac{A_K^2}{s_K^x s_K^y}$$

⑦ 增益值为

$$G_K = \begin{cases} A_K / s_K^x & \text{在 } f_K \text{ 处 } X_t \text{ 上的 } Y_t \text{ 的增益} \\ A_K / s_K^y & \text{在 } f_K \text{ 处 } X_t \text{ 上的 } Y_t \text{ 的增益} \end{cases}$$

⑧ 相位谱估计为

$$\Psi_K = \begin{cases} \arctan(Q_K / C_K) & \text{如果 } Q_K > 0, C_K > 0; Q_K < 0, C_K > 0 \\ \arctan(Q_K / C_K) + \pi & \text{如果 } Q_K > 0, C_K < 0 \\ \arctan(Q_K / C_K) - \pi & \text{如果 } Q_K < 0, C_K < 0 \end{cases}$$

(3) 数据窗口。

可以指定下面的频谱窗口。每个公式定义了窗口的上半部分。窗口的下半部分与上半部分对称。在所有公式中,  $p$  是跨度数除以 2 的整数部分。为简明扼要, 费热尔核公式被表达为

$$F_q(\theta) = \begin{cases} q & \theta = 0, \pm 2\pi, \pm 4\pi, \dots \\ \frac{1}{q} \left[ \frac{\sin(q\theta/2)}{\sin(\theta/2)} \right]^2 & \text{其他} \end{cases}$$

狄利克雷核公式被表达为

$$D_q(\theta) = \begin{cases} 2q+1 & \theta = 0, \pm 2\pi, \pm 4\pi, \dots \\ \frac{1}{q} \left[ \frac{\sin(2q+1)\theta/2}{\sin(\theta/2)} \right]^2 & \text{其他} \end{cases}$$

式中,  $q$  为任意的正实数。

① Tukey-Hamming 法。其计算权重公式为

$$W_k = 0.54D_p(2\pi f_k) + 0.23D_p\left(2\pi f_k + \frac{\pi}{p}\right) + 0.23D_p\left(2\pi f_k - \frac{\pi}{p}\right) \quad (k=0, \dots, p)$$

② Tukey 法。其计算权重公式为

$$W_k = 0.5D_p(2\pi f_k) + 0.25D_p\left(2\pi f_k + \frac{\pi}{p}\right) + 0.25D_p\left(2\pi f_k - \frac{\pi}{p}\right) \quad (k=0, \dots, p)$$

③ Parzen 法。其计算权重公式为

$$W_k = \frac{1}{p}[2 + \cos(2\pi f_k)][F_{p/2}(2\pi f_k)]^2 \quad (k=0, \dots, p)$$

④ Bartlett 法。其计算上半部分谱窗的权重为

$$W_k = F_p(2\pi f_k) \quad (k=0, \dots, p)$$

⑤ Daniell (Unit) 法。Daniell 窗口或矩形窗口。其计算谱窗形状的权重为

$$W_k = 1 \quad (k=0, \dots, p)$$

⑥ 无。不用平滑化处理。如果指定无, 则谱密度估计同周期图相同; 当跨度数为 1 时, 也如此, 即

$$W_{-p}, \dots, W_0, \dots, W_p$$

对于用户指定的权重, 如果权重数为奇数, 则中间的权重应用于已平滑的周期图值, 并且两边的权重应用于其前后值; 如果权重数为偶数(假定没有提供  $W_p$ ), 则在中间之后的权重应用于已平滑的周期图值。权重  $W_0$  一定是正数, 这是必须的。

## 17.7.2 频谱分析过程

- (1) 按【分析→预测→频谱分析】顺序单击菜单项, 打开【频谱图】对话框, 见图 17-36。
- (2) 在源变量框中选择一个或多个数值变量, 送入【变量】框中。
- (3) 在【频谱窗口】栏的下拉列表中选择平滑序列的滤波算法以便为获取谱密度估计做准备。可供选择的滤波法有:

- ① Tukey-Hamming 法。
- ② Tukey 法。
- ③ Parzen 法。
- ④ Bartlett 法。
- ⑤ Daniell(单位)法。
- ⑥ 无。不用作滤波处理。谱密度估计同周期图相同。

(4) 在【跨度】框中指定滤波的跨度值, 即横跨执行平滑的连续值的范围。通常使用奇整数。平滑谱密度图多使用大跨度, 较少使用小跨度。系统默认值为 5。

(5) 中心化变量的选择。

① 【中心(化)变量】。在计算频谱前, 校准序列使其有 0 均数(中心化), 并剔除同序列均数有关联的大量的项(剔除异常值)。因为谱分析时, 相应序列平均数频率为 0, 否则周期图没什么实际意义。因此先使数据以 0 为中心。



图 17-36 【频谱图】对话框

②【双变量分析】。如果选择两个或两个以上分析变量,可以选择该项,要求作【变量】框中第一个变量(因变量)与后面每个变量(自变量)的双变量谱分析。各个序列的单变量分析照样进行。

(6) 在【图】栏中选择输出的分析图。

【周期图】和【频谱密度】对单变量和双变量分析都有效,其他选项只对双变量分析有效。

①【周期图】。以频率或周期为横轴的非平滑的频谱振幅图(在对数标尺上绘制)。变异均匀地分布在所有波段象征“白噪声”。

②【平方一致性】。两序列增益值的乘积。

③【正交谱】。交叉周期图的虚部。它是两个时间序列异相频率分量相关的测度。分量是  $\pi/2$  弧度乘以异相。

④【交叉振幅】。余谱密度平方与正交谱平方之和的平方根,反映振幅的大小。

⑤【频谱密度】。已过滤去不规则变化的周期图。

⑥【余谱密度】。y 交叉周期图的实部,是两个时间序列同相频率分量相关的测度。

⑦【相位谱】。一个序列领先或滞后于其他序列的各频率分量的长度的测度。

⑧【增益】。用一个序列的谱密度除交叉振幅所得的商。两个序列中每一个都有其自己的增益值。它是在某一频率下的回归系数,同线性回归系数类似。

⑨【按频率】。所有图都由频率生成,频率的范围在频率 0(常数项或均数项)到频率 0.5(两个观察资料的周期项)之间。

⑩【按周期】。所有图都由周期生成,周期的范围在周期 2(两个观察资料的周期项)到周期等于观察值的数量(常数项或均数项)之间。周期在对数标尺上显示。

### 17.7.3 频谱分析实例

【例 7】数据文件 data17-05 中记录的是国际航线 1949 年 1 月至 1960 年 12 月间月度旅客总数(单位:千人),试用频谱分析法分析其是否有年度周期。

1) 打开数据文件后的操作步骤

(1) 按【分析→预测→频谱分析】顺序单击菜单项,打开如图 17-36 所示对话框。

(2) 选择 number 送入【变量】框中。在【图】栏中选择【频谱密度】。

(3) 单击【确定】按钮,执行运算。

2) 输出结果(见表 17-14 和图 17-33、图 17-34)

表 17-14 给出了模型的描述,从上至下依次是:模型名称(MOD\_12),分析类型(单变量),序列名 1(number),值范围(通过中心在 0 点处理),周期图平滑:谱窗口(Tukey-Hamming)、窗口跨度(5),权重值:  $W(-2)=2.233$ 、 $W(-1)=2.238$ 、 $W(0)=2.240$ 、 $W(1)=2.238$ 、 $W(2)=2.233$ 。

图 17-37 所示是周期图,图中显示的背景噪声中,有引人注目的连续的峰值,在小于 0.1 的最低频率处有最高的峰值,因此可以怀疑数据中包含一个年度的周期成分,年度成分的贡献组成了周期图。在时间序列中每个数据点表示一个月,因此一个年度周期对应于当前数据集中的周期 12。由于周期和频率互为倒数,周期 12 对应频率  $1/12$ (或 0.083),所以年度成分暗示在周期图中 0.083 处的一个峰值,它与正好低于 0.1 的频率处出现的峰值相一致。

图 17-38 所示是谱密度图,是经过消除背景噪声平滑后的周期图。残余峰值最好同谱密度函数一起分析,让潜在结构变得更加清楚独立。谱密度由 5 个明显的等间隔出现的峰值组成。最低频率峰值是在 0.08333 处。分析变量时间序列可以分解成 4 个主要(幅度较大)正弦或余弦成分。它们的周期即峰值点频率的倒数。

表 17-14 模型描述

模型描述		
模型名称	MOD_2	
分析类型	单变量	
序列名	1	numbwe
值范围	在零处通过居中减少	
周期图平滑	频谱窗口	Tukey-Hamming
	窗口跨度	5
	权重值	W(-2) 2.233
		W(-1) 2.238
		W(0) 2.240
		W(1) 2.238
		W(2) 2.233

正在应用来自 MOD\_2 的模型指定。

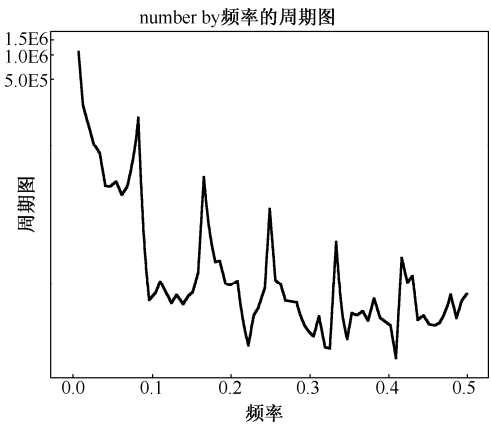


图 17-37 周期图

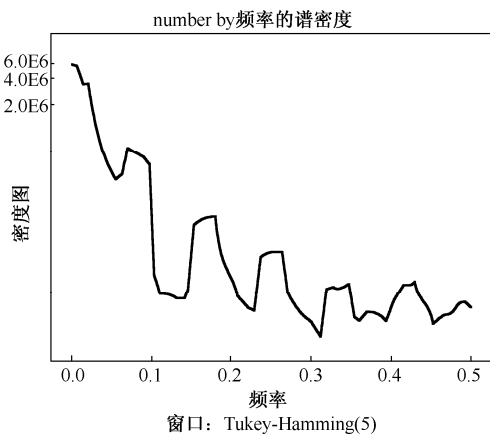


图 17-38 谱密度图

## 17.8 互 相 关

### 17.8.1 互相关概述

#### 1. 互相关的基本概念

ACF(自相关函数)和 PACF(偏自相关函数)是描述单个时间序列的重要工具。在很多场合下,需要考虑的时间序列不是一个,而是同时需要考虑多个时间序列之间的关系,如市场的货币供应量和股价变化之间的关系、某产品的广告投入和该产品市场占有率及销售量之间的关系。这时就需要考虑两个序列或多个序列之间的相互关系。为了和单序列分析(也称单变量时间序列分析)区分,将这里讨论的问题的模型称为多序列分析(或多元时间序列分析)。分析这种模型的工具是互相关函数。

所谓互相关函数(CCF)是指两个时间序列间的相关,即一个序列的观察值同另一个序列在不同的滞后和领先时的观察值之间的相关关系。互相关通常显示在图中,称为互相关图。互相关图可以帮助我们识别变量之间的关系。

#### 2. 互相关函数

设  $x$ 、 $y$  为长度为  $n$  的两个时间序列,则在滞后  $k$  处,  $x$  和  $y$  的互相关系数可用下式估计:

式中,

$$r_{xy}(k) = \frac{C_{xy}(k)}{S_x S_y}$$
$$C_{xy}(k) = \begin{cases} \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{x})(y_{t+k} - \bar{y}) & (k = 0, 1, 2, \dots) \\ \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \bar{y})(x_{t+k} - \bar{x}) & (k = -1, -2, \dots) \end{cases}$$

两个时间序列的标准差分别为

$$S_x = \sqrt{\frac{1}{n} \sum_{t=1}^n (x_t - \bar{x})^2}$$
$$S_y = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \bar{y})^2}$$

互相关函数关于  $k=0$  不对称。  
 $r_{xy}(k)$  的近似标准误为

$$se[r_{xy}(k)] = \sqrt{\frac{1}{n-|k|}} \quad (k = 0, \pm 1, \pm 2, \dots)$$

标准误是基于序列没有互相关和序列之一为白噪声假定的基础上计算的。

17.8.2 互相关过程

- (1) 该程序用来为正、负和 0 阶滞后绘制两个或多个序列互相关函数图。互相关程序只适用于时间序列数据。
  - (2) 按【分析→预测→互相关】顺序单击菜单项, 打开如图 17-39 所示【交叉相关性】对话框。
  - (3) 在源变量框中, 至少选择两个变量, 送入【变量】框中。
  - (4) 在【转换】栏中定义序列的转换方法。【自然对数转换】、【差分】、【季节性差分】3 个函数的说明可参阅 17.2.1 节中的相关内容。
- 在这些选项下面的【当前周期】后显示当前的周期。
- (5) 单击【选项】按钮, 打开【互相关性: 选项】对话框, 见图 17-40。

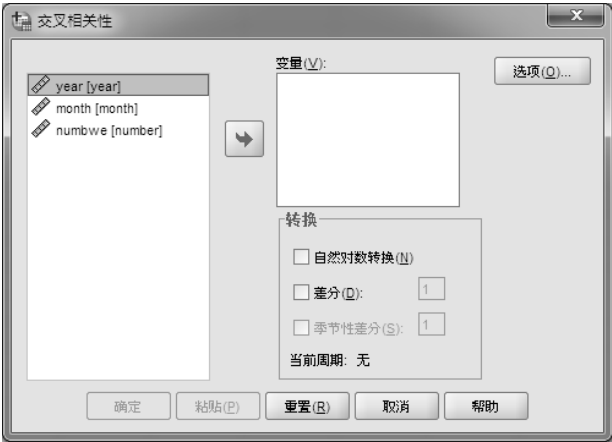


图 17-39 【交叉相关性】对话框



图 17-40 【互相关性: 选项】对话框

在【最大延迟数】框中输入互相关的最大滞后数，默认值为 7。如果数据中定义了季节，则只显示选项【在周期延迟处显示互相关】。

(6) 单击【继续】按钮，返回主对话框。单击【确定】按钮，系统立即执行命令。

17.8.3 互相关实例

【例 8】 数据文件 data17-06 中记录了 1989 年 1 月至 1998 年 12 月间某公司每月 3 种男、女服装产品的销售量情况，试分析这 10 年间男、女 3 种服装的销售量之间是否相关。

(1) 在数据编辑窗口中，打开数据文件 data17-06。按【分析→预测→互相关】顺序单击菜单项，打开如图 17-40 示对话框。

(2) 选择男装销售额变量 men 和女装销售额变量 women，送入【变量】框中。

(3) 其他保持系统默认选项，单击【确定】按钮，执行运算。

输出结果见表 17-15～表 17-17 和图 17-37。

表 17-15 所示是模型的描述。

表 17-16 所示是样品处理摘要。

表 17-15 模型描述

模型描述		
模型名称	MOD_1	
序列名	1	Sales of Men's Clothing
	2	Sales of Women's Clothing
转换	无	
非季节性差分		0
季节性差分		0
季节性期间的长度	无周期性	
滞后范围	从	-7
	至	7
显示并绘图	所有滞后	

正在应用来自 MOD\_1 的模型指定。

表 17-16 样品处理摘要

个案处理摘要		
序列长度		120
由于以下原因排除的个案数	用户缺失值	0
	系统缺失值	0
有效个案数		120
差分后可计算的零阶相关数		120

表 17-17 互相关系数表

交叉相关性		
序列对: 带有 Sales of Women's Clothing 的 Sales of Men's Clothing		
滞后	交叉相关	标准误差 <sup>a</sup>
-7	.159	.094
-6	.150	.094
-5	.211	.093
-4	.224	.093
-3	.271	.092
-2	.342	.092
-1	.374	.092
0	.802	.091
1	.134	.092
2	.114	.092
3	.125	.092
4	.209	.093
5	.163	.093
6	.124	.094
7	.178	.094

a. 基于以下假设：序列不具有交叉相关性，并且其中一个序列是白噪声。

表 17-17 所示是互相关系数的计算结果表，从左至右依次列出的是滞后、互相关系数值和互相关系数的标准误差。

图 17-41 所示是男、女服装销售量之间的互相关图，它用表 17-17 中滞后的值作为横坐标，用互相关系数值作为纵坐标，通过直方图的形式表现。最大互相关系数出现在滞后 0 处，为 0.802，显然，互相关系数并不关于滞后 0 处对称。滞后 0 处的相关同简单的两个变量间的皮尔

逊相关是一样的。说明两个变量之间存在线性正相关，也就是男装销售量大时，女装的销售量也在变大。图中平行于横轴的上、下两根横线分别是置信限的上、下限。

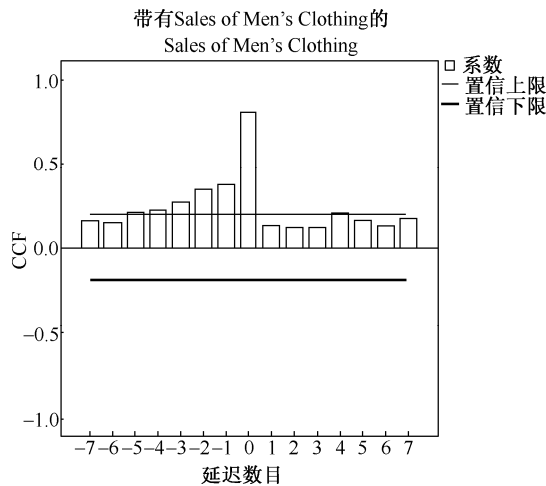


图 17-41 男女服装销售量之间的互相关图

## 习 题 17

1. 简述时间序列的基本概念。时间序列分析过程中有哪几种常用的方法？
2. 对数据用时间序列模型进行拟合处理前，应做哪些准备工作？
3. 在哪个过程中可进行缺失值的修补？修补缺失值的方法共有几种？
4. 在哪个过程中可定义时间变量？
5. 时间序列分析是建立在序列的平稳的条件上的，怎样判断序列是否平稳？
6. 为什么要建一个时间序列的新变量？在 SPSS 的哪个过程中建时间序列新变量？
7. 光盘中的数据文件 `data17-07.sav` 记录了一个邮购公司在 1989 年 1 月至 1998 年 12 月间男、女服装产品的销售量以及一些可能影响服装销售的宣传、服务方面的变量。试用学过的时间序列方法对其进行分析，并预测 1999 年 3 月的男装销售量。



# 第 18 章 生存分析

## 18.1 生存分析概述

### 18.1.1 生存分析与生存数据

生存分析广泛应用于生物医学、工业、社会科学、商业等领域，如肿瘤患者经过治疗后生存的时间、电子设备的寿命、罪犯假释的时间、婚姻的持续时间、保险人的索赔等。这类问题的数据特点是在研究结束时，所要研究的事件还没有发生或过早终止，使要收集的数据发生缺失。这样的数据称为生存数据。生存分析就是要处理、分析生存数据。

#### 1. 生存分析的类型与 SPSS 过程

生存分析方法可分为三类：非参数法、寿命表法和乘积极限法（用于估计生存率）；Log-rank 检验用于单因素预后分析；半参数模型即比例风险模型，用于辨认多协变量的预后因素；参数法一般也用作预后分析，如指数模型、Weibull 模型等。如果了解生存数据是否服从某特定分布，那么参数检验比非参数和半参数检验更有效。

在【分析】菜单下的【生存函数】子菜单中，提供了【寿命表】、【Kaplan-Meier】、【Cox 回归】、【Cox 依时协变量】（带时间相依性变量的生存分析）4 种生存分析方法，如图 18-1 所示。



图 18-1 【生存函数】子菜单

#### 2. 生存分析的数据

生存数据包括生存时间以及与其相关因素。生存数据有一个最重要的特点：在研究结束时在某些个体上还没有发生。所观测的含有这些事件的数据称为删失数据（Censored Data），也称截尾数据。如果生存数据中没有删失观测，则该生存数据称为完全数据。

按照删失数据发生的时间，删失数据可分为右删失、区间删失和左删失。例如，病人在研究期内被追踪观察某一事件的出现，如果在研究结束时刻，病人的该事件并未出现，则观察到的生存时间就是右删失数据；如果在研究期中间的某个时间区间中该事件出现了，则观察到的生存时间就是区间删失数据；如果在选进研究之前该事件已经出现了，则观察到的生存时间就是左删失数据。右删失类型包括单式删失和随机删失，常用的删失类型主要有 I 型和 II 型删失，它们均属于单式删失。动物试验、设备寿命研究中常遇到删失数据。

由于时间和费用受到限制，研究者常常不能等到所有动物死亡。事先确定截止观测的日期称为 I 型删失，又称定时删失；如果选择试验进行到有一固定数目的动物死亡为止，称为 II 型删失，又称定数删失。一般在数据的右上角标注“+”号表示是删失数据。

18.1.2 生存时间函数

生存时间测量某事件出现的时间。通常用下列 3 个函数来描述：生存函数、概率密度函数和危险率函数。它们在数学上是等价的，得出其中一个，可以推导出另两个。

生存函数也称累积生存率，记作  $S(t)$ ，它是指个体生存时间长于  $t$  的概率，即

$$S(t)=P(\text{个体生存时间长于 } t)$$

概率密度函数，又称密度函数，记作  $f(t)$ ， $f(t)$  的图形称为密度曲线，在任何时间区间内死亡的比例和死亡出现机会的峰值都可以在密度曲线上找出。函数表达式为

$$f(t)=\lim_{\Delta t \rightarrow 0} \frac{P(\text{个体在区间 } (t, t+\Delta t) \text{ 中死亡})}{\Delta t}$$

危险率函数，又称风险函数、瞬间死亡率、死亡强度、条件死亡率、分年龄死亡率、危险率，记作  $h(t)$ ，危险率函数是生存分析最基本的函数，即

$$h(t)=\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(\text{年龄是 } t \text{ 的个体在 } (t, t+\Delta t) \text{ 中死亡})$$

18.1.3 Kaplan-Meier 法

Kaplan-Meier 它由英国统计学家 Kaplan 和 Meier 于 1958 年提出，也称乘积限法。它适用于小样本或大样本未分组资料生存率的 Kaplan-Meier 法生存率估计及组间生存率比较。

1. 生存率的点估计

在时间  $t_i$  处的生存率估计为

$$\hat{S}(t)=\left(1-\frac{d_1}{n_0}\right)\left(1-\frac{d_2}{n_1}\right)\cdots\left(1-\frac{d_i}{n_{i-1}}\right) \quad (i=1, 2, \cdots, k)$$

式中， $n_{i-1}$ 、 $n_i$ 、 $d_i$  分别为活过时间  $t_{i-1}$  且未在  $t_{i-1}$  删失的观察对象数、期初例数(最初研究时的观察对象数)和死亡数。

2. 生存率的区间估计

Greenwood 生存标准误的近似计算公式为

$$SE\left[\hat{S}(t_i)\right]=\hat{S}(t_i)\sqrt{\sum_{j=1}^i \frac{d_j}{n_j(n_j-d_j)}}$$

在总体服从正态分布时，总体生存率的  $(1-\alpha)$  的置信区间为

$$\hat{S}(t_i) \pm Z_{\alpha/2} \cdot SE\left[\hat{S}(t_i)\right]$$

3. 组间的生存率比较

不同组之间的生存率比较简称组间的生存率比较，通常采用非参数的对数秩(Log rank)检验，其基本思想为，当  $H_0$ ：组间生存率相等为真时，根据  $t_i$  时点的死亡率，可以计算得到各组理论死亡率，由此可得

$$\chi^2=\frac{\left[\sum w_i(d_{gi}-T_{gi})\right]^2}{V_g}$$

式中,  $V_g$  为第  $g$  组理论数  $T_g$  的方差估计,  $V_g = \sum w_i^2 \frac{n_{gi}}{n_i} \left( 1 - \frac{n_{gi}}{n_i} \right) \left( \frac{n_i - d_i}{n_i - 1} \right) d_i$ 。  $w_i$  为权重, 在 Log rank 检验中,  $w_i = 1$ ; 在 Breslow 检验或 Wilcoxon 检验中,  $w_i = n_i$ ; 而在 Tarone-Ware 检验中,  $w_i = n_i^{1/2}$ 。 其中,  $n_i$  为时间  $t_i$  处对应的期初例数。

$\chi^2$  近似服从自由度为 (组数-1) 的  $\chi^2$  分布。

出现当前的  $\chi^2$  值及其更加极端值的概率不足 0.05 时, 拒绝原假设  $H_0$ 。

在作有序分类变量的多组间生存率的比较中, 如果 Log rank 检验组间生存率的差异有统计学意义, 则还可作趋势检验, 以此来进一步分析风险率是否有随分组等级的变化而变化的趋势。

### 18.1.4 Cox 回归模型

当众多的危险因素对生存时间有影响时, 应关心这其中哪些危险因素对生存时间有重要的影响, 也就是确认重要的预后因素 (预后因素是早已存在且与生存时间相关的因素)。 通过建立生存时间随危险因素变化的回归模型, 来确定这些对生存时间有影响的预后因素, 并根据危险因素在模型中的影响对生存率进行预测。 但是危险率往往难以估计, 所以不宜采用非参数或参数模型方法。 1972 年英国统计学家 D.R.Cox 提出了比例风险模型 (the Proportional Hazard Model, PHREG), 该模型可以很好地解决上述问题, 故又称 Cox 回归或 Cox 模型。 Cox 模型在表达形式上与参数模型相似, 但对各参数进行估计时又不依赖特定分布的假设, 所以也称半参数回归模型。 当生存时间是连续分布且预后变量间相互作用可被忽视时, 危险率  $h(t)$  为

$$h(t) = h_0(t) e^{(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)}$$

式中,  $h_0$  是基准的生存分布的危险率函数;  $\beta$  是回归系数;  $x$  是预后变量。 由于 Cox 模型假设, 每个预后变量的危险率在时间上正比于基础危险率 ( $h_0$ ), 从而无须计算 ( $h_0$ )。

Cox 模型除辨认预后因素外, 还可以确定预后指数或比率, 即求每个个体的  $\ln[h_i(t)/h_0(t)]$ 。

### 18.1.5 Cox 依时协变量回归模型

当 Cox 回归模型中的协变量对风险比例作用的强度随时间变化而变化时, 不再满足建立 Cox 回归模型的条件, 此时需改用 Cox 依时协变量回归模型, 也称非比例风险模型。

根据依存变量的取值和效应随时间变化的情况, 可将 Cox 依时协变量回归模型分成以下两种情形。

#### 1. 外在时间依存变量模型

当依存变量的取值不随时间改变, 但其效应 (RR) 随时间改变时, 称这种依存变量为外在时间依存变量。 此时的模型为

$$h(t, X) = h_0(t) e^{\beta X_E + r X_E t}$$

式中,  $h(t, X)$  为个体在协变量作用下, 在时点  $t$  的死亡率 (风险率);  $X$  为协变量向量;  $h_0(t)$  表示个体在时点  $t$  的基准风险率, 此时所有的协变量取值为 0;  $e^{\beta X_E + r X_E t}$  为医学上的相对风险度;  $X_E$  为时间依存变量, 其取值不随时间  $t$  改变;  $\beta$ 、 $r$  为回归系数。

#### 2. 内在时间依存变量模型

若依存变量的效应 (RR) 在不同时间点没有变化, 但其具体取值会随时间不同而改变, 则

称这种依存变量为内在时间依存变量。此时的模型为

$$h(t, X) = h_0(t)e^{\beta X_E(t)}$$

式中， $X_E$  表示变量取值在随时间变化而变化，其他变量的含义同上。

在上述情况下，需把可能随时间变化而变化的协变量定义成时间依存变量。当这样的协变量不止一个时，需用编程来进行分析。

## 18.2 寿命表分析

### 18.2.1 寿命表分析概述

寿命表(Life Table, LT)又称生命表。Mantel 和 Haznszel(1959 年)提出用寿命表的方法可以比较两种生存模式。寿命表分析方法是用来测定死亡率和描述群体生存现象的。一般说来，寿命表用来概括在特定的时期里特定人口的死亡情况，统称为人口寿命表。寿命表应用于患有某种疾病并且在一定时期受到跟踪研究的患者身上，对患者构造出的寿命表为临床寿命表。人口寿命表和临床寿命表在计算方法上是相似的，但所要求的数据来源不同。

寿命表用于大样本，并且对生存时间的分布不限，这是它的优点，所以它是目前广泛应用的一种非参数分析方法。在寿命表中，生存函数和生存率的估计依赖于寿命表中的所有区间。如果每个区间都很短，则区间个数很多，计算工作变得很繁重，不能体现其优点。尽管利用计算机分析使这项工作轻松简单，但输出的结果却十分冗长。用于寿命表的一个假定是总体在每个区间内各处有近似相等的生存概率。如果区间太长，这个假定可能受到破坏从而估计使不精确。

### 18.2.2 寿命表分析过程

#### 1. 寿命表分析基本过程

(1) 按【分析→生存函数→寿命表】顺序单击菜单项，打开【寿命表】对话框，见图 18-2。



图 18-2 【寿命表】主对话框

(2) 【时间】框。从左侧的变量框中选择“生存时间”变量进入该框。生存时间可以是任何时间单位，如果在生存变量中有负数，在分析过程中会将其剔除。

(3) 【显示时间间隔】栏。在该栏中确定时间的区间。默认单个值 0 作为时间区间的起点，用户在该选项的【到】框中输入所需要的时间区间的止点，在【步长】框中输入确定区间跨度的数值。例如，在【到】框中输入“200”，【步长】框中输入“20”，就表明时间区间的止点为 200 个时间单位，从 0 至 200 每 20 个时间单位为 1 个分组跨度。

(4) 【状态】框。选择状态变量进入该框中，该变量用来标定删失和非删失状态。单击【定义事件】按钮，打开【寿命表：为状态变量定义事件】对话框，见图 18-3。有两个选项：

①【单值】。默认单个变量值为 0。为状态变量选择一个值，则系统只对状态变量为该值观

测的生存时间进行分析，其他未选变量值的生存时间按删失值处理。例如，在状态变量中有 0、1、2、3 共 4 种变量值，如果在该框中输入“2”，则只对状态变量值为 2 的观测进行生存时间分析，而忽略其他 3 种变量值观测的生存时间分析。

②【值的范围】。指定状态变量值的范围。系统只对状态变量值在该范围内的观测的生存时间进行分析，其他值的生存时间按删失值处理。

(5)【因子】框。选择第一控制变量进入该框。单击【定义范围】按钮，打开【有效表格：定义因子】对话框，见图 18-4。不同处理方案导致不同的结果，选择第一控制变量进入【因子】框中将不同的方案结果分别显示。

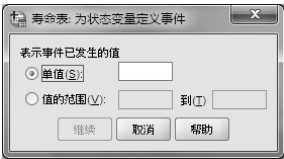


图 18-3 【寿命表：为状态变量定义事件】对话框

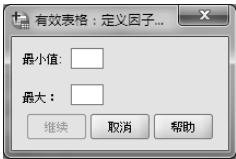


图 18-4 【定义控制变量范围】对话框

在【最小值】和【最大值】框中分别输入最小值和最大值，确定分析范围。不同的变量值代表不同的分层。其他未选变量值的生存时间按删失值处理，如果变量中有负值在分析过程中将被剔除。

(6)【按因子】框。选择第二控制变量进入该框，单击【定义范围】按钮，打开如图 18-4 所示的【定义控制变量范围】对话框，在【最小值】和【最大值】框中分别输入最小值和最大值，确定分析范围。第二分类变量中各分层将与第一分类变量中各分层相互结合，一一生成寿命表细分组。

2. 寿命表分析选择项

单击【寿命表】对话框中的【选项】按钮，打开【寿命表：选项】对话框，见图 18-5。

(1)【寿命表】选项。不选择该项，将不生成寿命表。

(2) 在【图】栏中选择生成的函数图形。

①【生存函数】。以线性刻度生成累积生存函数图。

②【取生存函数的对数】。以对数刻度生成累积生存函数图。

③【危险函数】。以线性刻度生成累积危险函数图。

④【密度】。生成密度函数图。

⑤【1 减去生存函数】。生成 1 减累积生存函数图。

(3) 在【比较第一个因子的水平】栏中，选择比较第一控制变量中各层间的显著性差异的方式，系统使用 Wilcoxon(Gehan)检验。如果有第二控制变量，先以第二控制变量的各层进行分组，再对第二控制变量中各分组中的第一控制变量中各层之间进行比较。

①【无】。不进行各分层的比较。

②【整体比较】。同时比较第一控制变量中各分层的差异。

③【两两比较】。两两比较第一控制变量中各层的差异。例如，在第一控制变量中有 3 个分层，将 1 对 2、2 对 3、1 对 3 分别进行比较，同时比较第一控制变量在各分层中的差异。

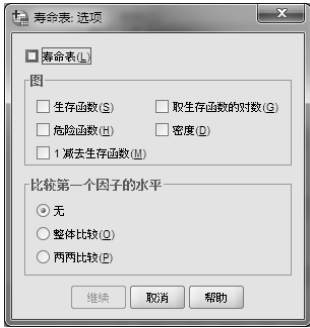


图 18-5 【寿命表：选项】对话框

### 18.2.3 寿命表分析实例

【例 1】有位科学工作者研究了饮食与肿瘤之间的关系，他将同种同龄的 90 只老鼠分成 3 组，在环境相同的情况下，分别给予低脂饮食(Low Fat)、饱和饮食(Saturated)和不饱和饮食(Unsaturated)，并对每只老鼠的脚趾注射等量的肿瘤细胞，观测这些老鼠 200 天。在这段时间内，有些老鼠偶然死亡且没有发现肿瘤，还有一些老鼠在观测结束时仍然没有肿瘤。以上资料源于《生存数据分析的统计方法》(Elisa T Lee 著，中国统计出版社)。

要求：作出不同喂养方式下的生存时间表，比较不同喂养方式下的生存时间是否有显著性差异，绘制各种函数图形。

#### 1) 数据文件

数据文件 data18-01.sav 中的变量有：ID 老鼠编号、FOOD 三种不同的喂养方式(编码与值标签是 1: low-fat, 2: saturated, 3: unsaturated)、STATUS 观测状态(编码与值标签为 0: died 已死亡, 1: censored 删失数据)、TIME 生存时间(天)。

#### 2) 数据处理

(1) 按【分析→生存函数→寿命表】顺序单击菜单项，打开【寿命表】对话框，见图 18-2。

(2) 从左侧的变量列表框中选择 TIME 变量，送入右侧的【时间】框。

(3) 在【显示时间间隔】栏中确定时间的区间。最后一个时间区间的开始点为“200”，步长(区间跨度)为“20”。

(4) 选择 STATUS 变量进入【状态】框。单击【定义事件】按钮，打开【寿命表：状态变量定义事件】对话框，见图 18-3，并在【单值】框中输入“0”。

(5) 选择 FOOD 变量进入【因子】框，作为第一控制变量。单击【定义范围】按钮，打开定义控制变量范围对话框，见图 18-4，在【最小值】和【最大值】框中分别输入“1”和“3”。

(6) 单击【寿命表】对话框中的【选项】按钮，打开【寿命表：选项】对话框，见图 18-5。选中【寿命表】复选项，选择【图】栏中的【生存函数】、【1 减去生存函数】选项，在【比较第一个因子的水平】栏中选择【两两比较】选项。

(7) 单击【确定】按钮，提交计算。

#### 3) 输出结果见表 18-1~表 18-5 和图 18-6。

表 18-1 所示为寿命表(表中的表头：年限表，汉化有误)，为能正确理解表中各列的内容，对表中各项解释如下：① 期初时间(时间区间)；② 期初记入数(进入时间区间的例数)；③ 期内退出数(活着退出的例数)(删失例数)；④ 历险数(暴露例数)；⑤ 期间终结数(时间区间中的死亡例数)；⑥ 终结比例(死亡率)；⑦ 生存比例(生存率)；⑧ 期末的累计生存比例(累积生存率)；⑨ 期末累计生存比例的标准误(累积生存率标准误差)；⑩ 概率密度(死亡概率)；⑪ 概率密度的标准误(死亡概率标准误)；⑫ 风险率(危险率函数)；⑬ 风险率的标准误(危险率标准误差)。

表 18-2 为中位生存时间。其中低脂肪食物老鼠中位生存时间为 197.93(月)。

用 Wilcoxon(Gehan)统计方法比较不同食物喂养老鼠导致癌症的生存时间。

表 18-3 所示为总体比较控制变量中不同水平的检验统计量。

表 18-4 所示为配对比较检验统计量。

表 18-5 所示为平均得分，其中包括总例数、未删失例数、删失例数、删失百分比、平均得分。

图 18-6(a)为生存函数图，图 18-6(b)为 1-生存函数图。

表 18-1 低脂肪食物的老鼠寿命表

年限表													
一阶段控制	期初时间	期初记入人数	期内退出数	历险数	期间终结数	终结比例	生存比例	期末的累积生存比例	期末的累积生存比例的标准误差	概率密度的标准误差	风险率	风险率的标准误差	
food 食物分类	饱和饮食	0	30	0	30.000	0	.00	1.00	1.00	.00	.000	.000	.00
	20	30	0	30.000	0	.00	1.00	1.00	.00	.000	.000	.00	.00
	40	30	0	30.000	4	.13	.87	.87	.06	.007	.003	.01	.00
	60	26	0	26.000	3	.12	.88	.77	.08	.005	.003	.01	.00
	80	23	0	23.000	6	.26	.74	.57	.09	.010	.004	.02	.01
	100	17	0	17.000	4	.24	.76	.43	.09	.007	.003	.01	.01
	120	13	0	13.000	3	.23	.77	.33	.09	.005	.003	.01	.01
	140	10	0	10.000	2	.20	.80	.27	.08	.003	.002	.01	.01
	160	8	1	7.500	1	.13	.87	.23	.08	.002	.002	.01	.01
	180	6	0	6.000	0	.00	1.00	.23	.08	.000	.000	.00	.00
	200	6	6	3.000	0	.00	1.00	.23	.08	.000	.000	.00	.00
	非饱和饮食	0	30	0	30.000	0	.00	1.00	1.00	.00	.000	.000	.00
	20	30	0	30.000	0	.00	1.00	1.00	.00	.000	.000	.00	.00
	40	30	0	30.000	0	.00	1.00	1.00	.00	.000	.000	.00	.00
	60	30	0	30.000	12	.40	.60	.60	.09	.020	.004	.03	.01
	80	18	0	18.000	5	.28	.72	.43	.09	.008	.003	.02	.01
	100	13	0	13.000	7	.54	.46	.20	.07	.012	.004	.04	.01
	120	6	0	6.000	1	.17	.83	.17	.07	.002	.002	.01	.01
	140	5	0	5.000	2	.40	.60	.10	.05	.003	.002	.03	.02
	160	3	0	3.000	3	1.00	.00	.00	.00	.005	.003	.10	.00
	低脂肪	0	30	0	30.000	0	.00	1.00	1.00	.00	.000	.000	.00
	20	30	0	30.000	0	.00	1.00	1.00	.00	.000	.000	.00	.00
	40	30	0	30.000	2	.07	.93	.93	.05	.003	.002	.00	.00
	60	28	0	28.000	4	.14	.86	.80	.07	.007	.003	.01	.00
	80	24	0	24.000	3	.13	.88	.70	.08	.005	.003	.01	.00
	100	21	0	21.000	1	.05	.95	.67	.09	.002	.002	.00	.00
	120	20	0	20.000	0	.00	1.00	.67	.09	.000	.000	.00	.00
	140	20	1	19.500	2	.10	.90	.60	.09	.003	.002	.01	.00
	160	17	0	17.000	1	.06	.94	.56	.09	.002	.002	.00	.00
	180	16	0	16.000	2	.13	.88	.49	.09	.004	.002	.01	.00
	200	14	14	7.000	0	.00	1.00	.49	.09	.000	.000	.00	.00

表 18-2 中位生存时间

中位数生存时间	
一阶段控制	中位数时间
食物分类 low-fat	197.93
saturated	110.00
unsaturated	92.00

表 18-4 配对比较检验统计量

成对比较 <sup>a</sup>				
(I) food	(J) food	Wilcoxon (Gehan) 统计量	df	Sig.
1	2	3.676	1	.055
	3	11.913	1	.001
2	1	3.676	1	.055
	3	2.532	1	.112
3	1	11.913	1	.001
	2	2.532	1	.112

a. 比较是精确的。

表 18-3 总体检验统计量

整体比较 <sup>a</sup>		
Wilcoxon (Gehan) 统计量	df	Sig.
12.058	2	.002

a. 比较是精确的。

表 18-5 平均得分

平均分					
比较组	总数	未审查	已审查	已审查的百分比	平均分
1 对比 2	1	30	15	50.0%	8.400
	2	30	23	77.0%	-8.400
1 对比 3	1	30	15	50.0%	15.400
	3	30	30	100.0%	-15.400
2 对比 3	2	30	23	77.0%	7.167
	3	30	30	100.0%	-7.167
1 对比 3	1	30	15	50.0%	15.400
	2	30	23	77.0%	7.167
2 对比 3	2	30	30	100.0%	-7.167
	3	30	30	100.0%	-7.167

整体比较

由于选择的观测较少，所以不太适合用寿命表分析方法，但从寿命表的终结比例即“死亡率”的数据可以看出，60~100 天内喂养低脂肪食物患肿瘤死亡率较高。从表 18-3 中可见，用 3 种不同食物喂养，老鼠患癌症后所生存的时间经过 Wilcoxon (Gehan) 检验，存在显著性差异 ( $p = 0.002 < 0.05$ )。在进行组间比较时，得到低脂 (low-fat) 食物和饱和 (saturated) 食物之间的生存时间有显著性差异，检验统计量为 11.913，自由度为 1，概率为 0.0006 ( $p < 0.05$ )。由图 18-6(a) 可见，随着生存时间的延长，累积生存率在下降，3 条曲线明显不重叠，可以直观地看出喂养不同食物的老鼠的生存时间有所不同，低脂肪组的生存时间最长 (累积生存率最高)，其次为饱和和饮食组，而不饱和饮食组的生存时间最短。图 18-6(b) 所示图形上下翻转 180° 后与图 (a) 是完全一样的，因而结论也是一致的。

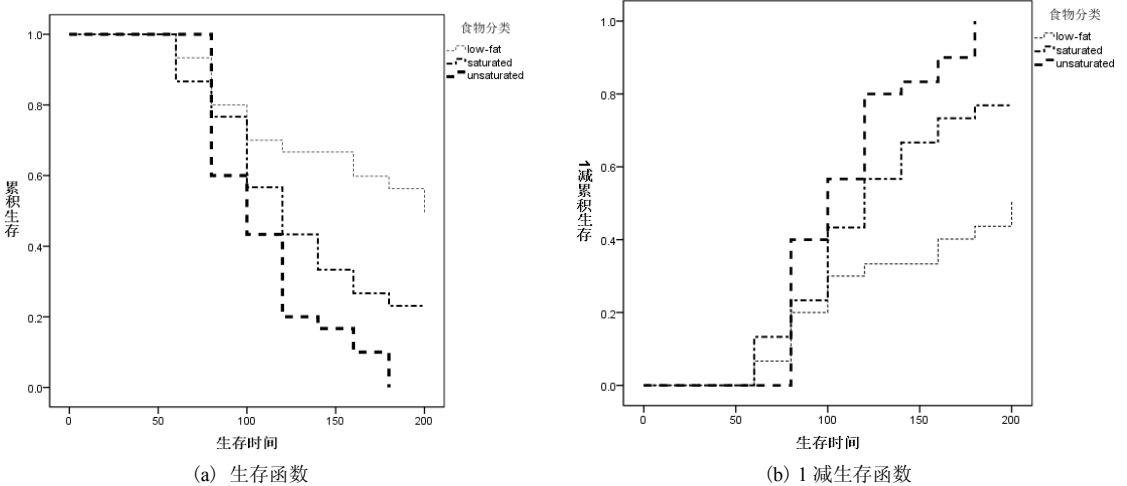


图 18-6 生存图形

## 18.3 Kaplan-Meier 分析

### 18.3.1 Kaplan-Meier 分析概述

对于 Kaplan 和 Meier (1958 年) 所提出的估计生存函数的乘积限 (Product-Limit, PL) 方法，很多人也把它称为寿命表估计，二者的差别是：PL 估计是基于一个个的数据，而寿命表估计基于按区间分组数据。PL 估计可看成是寿命表估计的特殊情形。

时间变量应是数值型。状态变量可以是二分变量或多分类变量，发生的事件可以用一个正数值表示或用某个范围的连续数值表示。

寿命表假设事件发生的概率仅依赖于时间。

### 18.3.2 Kaplan-Meier 分析过程

#### 1. Kaplan-Meier 分析基本过程

(1) 按【分析→生存函数→Kaplan-Meier】顺序单击菜单项，打开【Kaplan-Meier】主对话框，见图 18-7。

(2) 【时间】框。从左侧的变量框中选择生存时间变量进入该框，生存时间可以是任何时间单位，如果在生存变量中有负数，系统在分析过程中将其剔除。



(3) 【状态】框。选择标定删失和非删失的状态变量进入该框。单击【定义事件】按钮，打开【Kaplan-Meier: 定义状态变量发生事件】对话框，见图 18-8。在该对话框中选择要分析的状态，系统只分析选定的状态下的生存时间数据，其余按删失值处理。



图 18-7 【Kaplan-Meier】主对话框



图 18-8 【Kaplan-Meier: 定义状态变量发生事件】

① 【单值】。例如，在状态变量中有 0、1、2、3 共 4 种变量值，如果在该框中输入“2”，则只对状态变量值为“2”的观测进行生存时间进行分析。

② 【值的范围】。指定状态变量值范围。例如，在状态变量中有 0、1、2、3 共 4 种值，输入值为“1”和“3”，只对状态值为“1”、“2”、“3”的生存时间进行分析。

③ 【值的列表】。变量值列表。例如，在状态变量中有 0、1、2、3 共 4 种变量值，如果输入“1”和“3”，只分析状态值为“1”和“3”的生存时间。

(4) 【因子】框。选择控制变量进入该框。用短字符型或数值型的变量值代表不同水平。

(5) 【层】框。选择分层变量进入本框，即在控制变量中不同的处理方案内进行分层。该变量的值代表不同的分层。变量可以是短字符型也可以是数值型。

(6) 【标注个案】框。选择标识观测的变量进入该框，SPSS 将以列表方式用该变量值标出所有的观测。该变量可以是字符型，其值可以是小于等于 20 个字母的字符串。

## 2. 选择比较控制因素的统计方法

选择了控制变量后，可以比较各个不同水平是否具有显著性差异。单击【Kaplan-Meier】对话框中的【比较因子】按钮，打开【Kaplan-Meier: 比较因子水平】对话框，见图 18-9。

(1) 选择统计方法。

① 【对数秩】。即 Mantel-Haenszel 检验，又称时序检验，对所有的死亡时间赋予相等的权重，比较生存分布是否相同，它对于后期差别较为敏感。

② 【Breslow】。对较早死亡时间赋予较大的权重，所以对于早期差别较为敏感。

③ 【Tarone-Ware】。比较生存分布是否相同，当两个危险率函数曲线或生存曲线有交叉时，可以考虑使用 Tarone-Ware 检验。

(2) 选择比较的方式。

① 【因子水平的线性趋势】。如果因子水平有自然顺序(如病情的早期、中期、晚期)时，选中该复选项，作趋势检验。



图 18-9 【Kaplan-Meier: 比较因子水平】对话框

- ②【在层上比较所有因子水平】。合并比较所有因子水平下的生存时间，不进行配对比较。
- ③【对于每一层】。如果选择了分层变量，在每层比较不同因子水平下的生存时间。
- ④【在层上成对比较因子水平】。以不同的配对方式比较每一对因子水平下的生存时间，如果选择了趋势检验，这种方法不能使用。
- ⑤【为每层成对比较因子水平】。如果选择了分层变量，在每层以不同的配对方式比较每一对因子水平下的生存时间。但选择了趋势检验，这种方法也不能使用。

3. 保存新的统计量

将运算中新的统计量保存到数据窗中，单击【Kaplan-Meier】对话框中的【保存】按钮，打开【Kaplan-Meier: 保存新变量】对话框，见图 18-10。

- ①【生存函数】。保存累积生存概率估测值，如果没有指定变量名，自动生成前缀带有“sur”的变量名，如“sur\_1”，“sur\_2”等。
- ②【生存函数的标准误】。保存累积生存概率的标准误，如果没有指定变量名，自动生成前缀带有“se”的变量名，如“se\_1”，“se\_2”等。
- ③【危险函数】。保存累积危险函数估测值，如果没有指定变量名，自动生成前缀带有“haz”的变量名，如“haz\_1”，“haz\_2”等。
- ④【累积事件】。保存发生事件的累积频率，如果没有指定变量名，自动生成前缀带有“cum”的变量名，如“cum\_1”，“cum\_2”等。

4. Kaplan-Meier 分析选择项

用户根据需要选择一些统计量和图形。单击【Kaplan-Meier】对话框中的【选项】按钮，打开【Kaplan-Meier: 选项】对话框，见图 18-11。

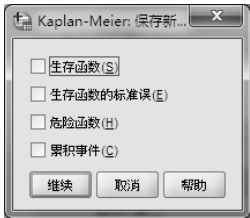


图 18-10 【Kaplan-Meier: 保存新变量】对话框

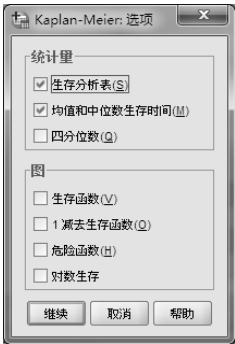


图 18-11 【Kaplan-Meier: 选项】对话框

(1) 【统计量】栏。

- ①【生存分析表】。生成一个简化的寿命表，它只包括乘积限寿命表、标准误、累积频数、风险例数。如果清除该复选项，将不生成寿命表，这样可以压缩输出的篇幅。
- ②【均值和中位数生存时间】。计算生存时间的均数、中位数及其标准误和置信区间。
- ③【四分位数】。输出结果显示生存时间的 25、50 和 75 分位数，以及它们的标准误。

(2) 在【图】栏中选择生成的函数图形。

- ①【生存函数】。将生成线性刻度的累积生存函数图。
- ②【1 减去生存函数】。生成 1 减累积生存函数图。
- ③【危险函数】。生成线性刻度的累积危险函数图。
- ④【对数生存】。生成对数刻度的累积生存函数图。

18.3.3 Kaplan-Meier 分析实例

【例 2】某医院对 58 例肾上腺样瘤患者在不同治疗研究中得到的数据，资料源于《生存数据分析的统计方法》(ELISA T. LEE 著，中国统计出版社)，数据文件为 data18-02.sav。

要求显示生存时间的均数和中位数，以及 25 分位数、50 分位数和 75 分位数；检验在切除肾脏条件下两种治疗方案的结果是否具有显著性差异。

1) 数据

数据文件 data18-02 中的变量、变量标签、值、值标签为：id(患者编号)、sex(性别：1，男；2，女)；k(肾切除情况：0，未切；1，切除)、tre(治疗方案：1，化学与免疫疗法结合；2，其他方法)、time(生存时间：-99，未知)、sta(观测的状态：0，删失数据；1，已死亡；9，未知)。

2) 操作步骤

- (1) 按【分析→生存函数→Kaplan-Meier】顺序单击菜单项，打开如图 18-7 所示对话框。
  - (2) 从左侧的变量框中选择 time 变量，送入【时间】框中。
  - (3) 选择 sta 变量进入【状态】框中。单击【定义事件】按钮，打开【Kaplan-Meier：定义状态变量发生事件】对话框，见图 18-8，并在【单值】框中输入“1”。
  - (4) 选择 tre 变量进入【因子】框，作为控制变量。
  - (5) 选择 k 变量进入【层】框，作为分层变量。
  - (6) 单击【比较因子】按钮，打开【Kaplan-Meier：比较因子水平】对话框，见图 18-9。选中【对数秩】选项，同时选中【为每层成对比较因子水平】选项。
  - (7) 单击【选项】按钮，打开【Kaplan-Meier：选项】对话框，见图 18-11。选中【均值和中位数生存时间】和【四分位数】复选项。
  - (8) 单击【确定】按钮，提交计算。
- 3) 输出结果(见表 18-6~表 18-9)

表 18-6 观测删失情况

个案处理摘要					
肾切除情况 治疗方案		总数	事件数	删失	
				N	百分比
未切	化学与免疫法结合	7	7	0	0.0%
	其他方法	3	3	0	0.0%
	整体	10	10	0	0.0%
切除	化学与免疫法结合	29	25	4	13.8%
	其他方法	17	12	5	29.4%
	整体	46	37	9	19.6%
整体	整体	56	47	9	16.1%

表 18-7 生存时间的平均值和中位数

生存表的均值和中位数									
肾切除情况 治疗方案		均值 <sup>a</sup>				中位数			
		估计	标准误	95% 置信区间		估计	标准误	95% 置信区间	
				下限	上限			下限	上限
未切	化学与免疫法结合	12.571	2.034	8.585	16.558	12.000	3.928	4.301	19.699
	其他方法	8.000	.000	8.000	8.000	8.000	.	.	.
	整体	11.200	1.555	8.152	14.248	8.000	.949	6.141	9.859
切除	化学与免疫法结合	46.217	7.154	32.194	60.240	36.000	7.908	20.500	51.500
	其他方法	52.392	18.232	16.657	88.128	20.000	4.749	10.692	29.308
	整体	47.414	7.698	32.326	62.503	30.000	6.982	16.316	43.684
整体	整体	40.825	6.579	27.929	53.720	20.000	3.606	12.932	27.068

a. 如果估计值已删失，那么它将限制为最长的生存时间。

表 18-8 生存时间的四分位数

		百分位数					
肾切除情况	治疗方案	25.0%		50.0%		75.0%	
		估计	标准误	估计	标准误	估计	标准误
未切	化学与免疫法结合	17.000	2.315	12.000	3.928	8.000	1.793
	其他方法	8.000	.	8.000	.	8.000	.
	整体	15.000	3.795	8.000	.949	8.000	.791
切除	化学与免疫法结合	72.000	16.537	36.000	7.908	14.000	3.404
	其他方法	40.000	8.277	20.000	4.749	16.000	2.627
	整体	68.000	12.163	30.000	6.982	14.000	2.962
整体	整体	52.000	14.250	20.000	3.606	10.000	1.614

表 18-6 所示为 KM 分析过程中观测删失情况，KM 分析过程中将变量中的负数或缺失值剔除。Total N 总数、N of Events 未删失的例数、Censored N 删失数、Censored Percent 删失的百分比。

表 18-7 和表 18-8 所示为不同分层及不同处理情况生存描述性统计量。表 18-7 所示为生存时间的平均值和中位数以及它们 95% 的置信区间。表 18-8 所示为生存时间的四分位数，即 25%、50%、75% 的数值。

表 18-9 所示为 Log Rank 检验统计量，在分层变量为 0 值时，对控制变量不同的水平作时序检验 (Log Rank)。

表 18-9 Log Rank 检验统计量

成对比较						
			化学与免疫法结合		其他方法	
			卡方	Sig.	卡方	Sig.
Log Rank (Mantel-Cox)	肾切除情况	治疗方案				
		未切	化学与免疫法结合			2.440
		其他方法	2.440	.118		
	切除	化学与免疫法结合			.110	.741
其他方法		.110	.741			

统计结果表明，对 58 名肾上腺样瘤的治疗中，无论患者的肾脏切除或不切除，化学与免疫结合的疗法同其他疗法，在延长患者生存时间上没有显著性差别。在肾脏切除的情况下，Log Rank 检验统计量为 2.44 ( $p > 0.05$ )；在肾脏未切除的情况下，Log Rank 检验统计量为 0.11 ( $p > 0.05$ )。

18.4 Cox 回归风险比例模型分析

18.4.1 Cox 回归分析概述

在 Cox 回归模型中，生存时间或恢复时间常作因变量，而与生存时间有关的一组变量作为自变量，即预后变量或协变量。

时间变量应是数值型。状态变量可以是分类或连续型变量。如果是分类变量，应是哑变量或用指示编码。分层变量为分类变量，可用整数或短字符串编码。自变量(协变量)可以是分类型或连续型变量。预后变量可以是连续变量或离散变量。连续型自变量可以直接用在方程里；若是离散型变量，必须编码成指示变量才能参与分析。指示变量可以在定义分类协变量对话框重新编码。关于指示变量的编码方式见第 11 章。

在拟合 Cox 模型之前，可以通过计算变量之间的相关系数来查明与因变量显著相关的变

量,对数据的质量进行检查,然后结合专业知识拟合模型。应注意没有进入模型中的因素不一定是无关的因子,进入模型中的因子也不一定就是相关因子。

比例风险假设为,两组被试对象在任何时间点上发生事件的风险比例是恒定的。

## 18.4.2 Cox 回归分析过程

### 1. Cox 回归分析基本过程

(1) 按【分析→生存函数→Cox 回归】顺序单击菜单项,打开【Cox 回归】主对话框,见图 18-12。

(2) 【时间】框。从左侧的变量列表中选择生存时间变量进入该框,生存的时间可以是任何时间单位的连续型变量。在分析中自动剔除生存变量值为负数的观测(个案)。

(3) 【状态】框。选择标定删失和非删失状态的状态变量进入该框中。单击【定义事件】按钮,打开【定义状态变量发生事件】对话框,见图 18-8,选择要分析的状态,具体方法见 18.3.2 节。

(4) 【协变量】框。从变量列表框中选定一个或多个协变量进入该框,协变量可以是连续型或分类型变量。



图 18-12 【Cox 回归】主对话框

通过使用【上一张(组)】与【下一张(组)】按钮,可指定不同的协变量组,单击【下一张】按钮进入下一个协变量组,单击【上一张】按钮退回上一个协变量组。

如果考虑协变量间的交互作用,在变量列表中选择有交互作用的变量,单击【a\*b】按钮,形成交互作用项进入【协变量】框。

(5) 【方法】下拉列表。在该下拉列表中选择协变量进入回归模型的方式,共 7 种。

① 【进入】。强行进入法,同一组中的协变量,一次性地全部进入回归方程。

② 【向前: 条件】。变量经过条件似然检验确定是否进入模型的向前选择法。

③ 【向前: LR】。变量经过似然率检验确定是否进入模型的向前选择法。

④ 【向前: Wald】。变量经过沃德检验确定是否进入模型的向前选择法。

⑤ 【向后: 条件】。变量经过条件似然检验确定是否从模型中剔除的向后选择法。

⑥ 【向后: LR】。变量经过似然比检验确定是否从模型中剔除的向后消去法。

⑦ 【向后: Wald】。变量经过沃德检验确定是否从模型中剔除的向后消去法。

一般来说,使用向后消去法可以减少漏掉潜在的有价值的预测因子。如果至少有一个协变量进入模型,可以使用向前选择法。

(6) 【层】框。选定分层变量进入该框,SPSS 根据分层变量将数据细分组,然后在每个分组数据的基础上生成各自的风险函数。分层变量应是分类变量。

### 2. 分类变量的编码

在主对话框中单击【分类】按钮,打开【Cox 回归: 定义分类协变量】对话框,见图 18-13。对于数值型的分类变量需要在该对话框中重新编码,新的编码变量名后标注“Cat”。



图 18-13 【Cox 回归：定义分类协变量】对话框

一类都与该类前面的各类的平均效应相比较。又称反赫尔默特对比。

④【Helmert】(赫尔默特对比)。除最后一类外,预测变量的每类与后面各类的平均效应相比较。

⑤【重复】。除第一类外,预测变量的每个分类都与它前面的分类比较。

⑥【多项式】(正交多项式对比)。只能用于数值型分类变量,且假定各类间有相等的空间。

⑦【指示符】(指示对比)。指明类代表信息的有无,可选第一个或最后一个类别作为参考类。

⑧【参考类别】。离差对比、简单对比和指示对比中,用户可以去除默认的参照分类,可以选择第一个或最后一个分类作为默认分类。

完成选择后,单击【更改】按钮,确定这些设置。

### 3. 生成图形

单击【Cox 回归】对话框中的【绘图】按钮,打开【Cox 回归：图】对话框,见图 18-14。在该对话框中,用户可以获得以下图形(如果有时间相依性协变量,则不能生成图形)。

(1)【图类型】栏。

①【生存函数】。生成线性刻度的累积生存函数图形。

②【危险函数】。生成线性刻度的累积危险函数图形。

③【负对数累积生存函数的对数】。生成经过  $\ln(-\ln)$  转换之后的累积生存估计值的图形。

④【1 减去生存函数】。生成 1 减累积生存函数图。

(2)【协变量值的位置】框。在默认状态下,以模型中对比变量和协变量的均值绘制函数图形,也就是在该框中单击【均值】选项,再单击【更改】按钮。如果以对比变量和协变量的其他数值绘制函数图形,选中该框中的一个或多个协变量,然后在【更改值】框中选择【值】选项,并在其后参数框中输入数值,最后单击【更改】按钮,SPSS 将根据用户指定的协变量值,绘制其危险函数和生存函数。

(1)【协变量】框中为所有在主对话框中选定的协变量。从中选择要编码的数值型分类自变量送入【分类协变量】框。

(2)在【分类协变量】框中选择一个变量,在【更改对比】栏中选择一个对比类型和对比类。

①【偏差】。预测变量中每个分类效应与总效应比较。

②【简单】。预测变量的每类与参照类比较。可选择第一类或最后一类作为参考类。

③【差值】。除第一类外,预测变量的每

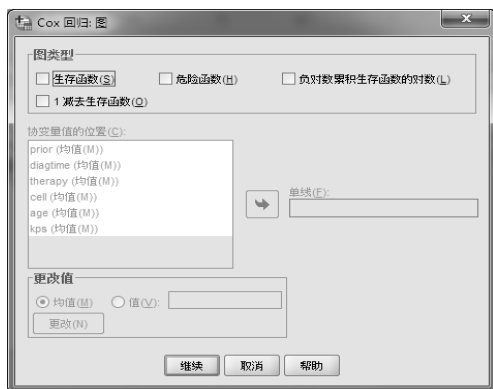


图 18-14 【Cox 回归：图】对话框

(3) **【单线】**框。选择一个分类协变量进入该框，系统按变量值将数据分成两个或多个分组，对各分组分别绘制函数图。如果指定了层变量，则每层绘制一个图。

#### 4. 保存新的统计量

在主对话框中单击**【保存】**按钮，打开如图 18-15 所示的**【Cox 回归：保存新变量】**对话框。选择要保存在数据窗中的分析结果，作为新变量有生存变量和诊断变量。

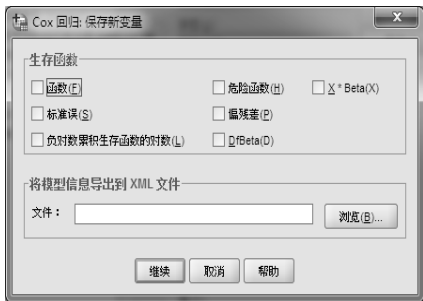


图 18-15 **【Cox 回归：保存新变量】**对话框

(1) **【生存函数】**栏指定生成的生存变量。

① **【函数】**。保存生存函数估测值，自动生成的变量名前缀为“sur”，如 sur\_1、sur\_2 等。

② **【标准误】**。生存函数估测值的标准误。自动生成的变量名前缀为“se”。

③ **【负对数累积生存函数的对数】**。经对数-对数转换的生存函数估测值。新变量名前缀为“lml”。

④ **【危险函数】**。累积危险函数估测值。自动生成的变量名前缀为“haz”。

⑤ **【偏残差】**。生成对生存时间的偏残差，用以检验比例危险的假设，SPSS 为最终模型中每个协变量保存一个偏残差变量。在模型中至少含有一个协变量才能生成偏残差。自动生成的变量名前缀为“pr”，如 pr1\_1、pr1\_2、pr2\_1、pr2\_2 等。

⑥ **【DfBeta(s)】**。每个观测从模型拟合中被剔除时，标准化回归系数的变化量。模型中至少含有一个协变量才能生成标准化回归系数变化量变量。新变量名前缀为“dfb”。

⑦ **【X\*Beta】**。线性预测因素得分。它是平均中心协变量值与其相对应的每个观测参数估计值的乘积和。新变量名前缀为“xbe”。

(2) 将模型信息导出到 XML 文件。

在**【文件】**框中输入需要保存文件的路径和名称，可将参数估计值导出到 XML 格式的文件中。在需要应用该模型信息对其他数据文件进行评分时，用户可以使用该模型文件。

#### 5. Cox 回归分析选项

在主对话框中单击**【选项】**按钮，打开**【Cox 回归：选项】**对话框，选择统计和输出方式，见图 18-16。

(1) **【模型统计量】**栏。

① **【CI 用于 exp(B)】**。设置相对危险估计值的置信区间，常用的有 90%、95%和 99%。

② **【估计值的相关性】**。显示回归系数估计值的相关系数矩阵。

③ **【显示模型信息】**栏。对当前模型显示对数似然统计量、似然比统计量和总体卡方值。对模型中的变量，显示参数估计值及其标准误，Wald 统计量。对已剔除出模型的变量，显示记分检验统计量和残差卡方值。

• **【在每个步骤中】**。在逐步回归的每一步显示上述全部统计量。

• **【在最后一个步骤中】**。显示逐步回归最后一步进入模型的协变量和最后模型的统计量。

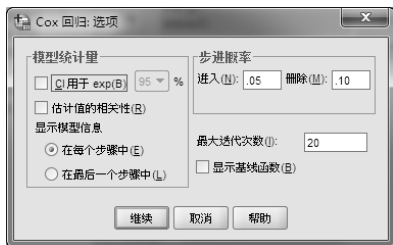


图 18-16 **【Cox 回归：选项】**对话框

(2)【步进概率】栏。如果选择了逐步回归法，还应在【进入】和【删除】框中指定协变量进入或剔除出模型的概率，默认的概率分别为“0.05”和“0.10”。注意，进入概率值应该小于剔除概率值，否则模型中将没有变量。

(3)【最大迭代次数】框。为模型指定最大迭代次数。用 Newton-Raphson 方法计算参数估计值时，如果达到最大迭代次数，则迭代过程将停止。

(4)【显示基线函数】。生成基准危险函数、协变量均值生存和危险函数表。若有分层变量，则每层生成独立表格。若指定了时间相依性协变量，不能激活该选项。

6. 自举法

单击【Bootstrap】按钮，打开【Bootstrap】对话框，见图 18-17。



图 18-17 Bootstrap 对话框

Bootstrap 法是一种从原样本中通过有放回重复抽取与原样本量相等样本的一种抽样方法。

仅当选择【执行 Bootstrap】选项时，对话框中的其他选项才处于激活状态。

(1) 设定样本数。在【样本数】框中可输入一个正整数来设定所需生成的样本数。系统默认值为“1000”。

(2) 设定随机种子。选择【设置 Mersenne Twister 种子】选项，可在【种子】框中输入一个正整数来设定随机种子。

(3) 设定置信区间。系统默认的置信区间为 95%，可在【水平 (%)】框中自定义数值对默认的置信区间进行修改。

置信区间的类型可以选择①【百分位】，也可以选择②【偏差修正加速(BCa)】选项。

(4) 抽样。有两种抽样方式可供选择。

- ①【简单】。采用简单随机抽样。
- ②【分层】。采用分层抽样。如果选择本项，则在其下方的变量框中选择一个变量作为分层变量，并将其移入【分层变量】框中。

使用本框中的【Bootstrap 方法】可以导出稳健的标准误估计值，并能为均值、中位数、比例、几率比、相关系数或回归系数等估计值计算置信区间。此外，它还可用来构建假设检验。当参数估计方法的假设存疑时(如拟合较小样本的异方差残差回归模型)，无法执行参数推论或需要非常复杂的标准误计算公式(如为中位数、四分位数和其他百分位数计算置信区间)时，Bootstrap 是最好的备选项。

18.4.3 Cox 回归分析实例

【例 3】数据文件 data18-03.sav 中是一组 137 位肺癌患者生存时间的数据。该数据来自《SAS/STAT guide for personal computers》，用 Cox 回归模型辨认预测因素。

1. 需要说明的变量：

diagtime 诊断到治疗的时间，time 生存时间，prior 治疗前处理(0：经过处理；1：未经处理)，therapy 治疗方案(1：标准方法；2：实验方法)，status 病人状态(0：死亡；1 删失数据)，



cell 肺癌细胞组织学分类(1: 鳞癌细胞; 2: 小细胞; 3: 腺癌细胞; 4: 大细胞), kps 判断标准( $\leq 30$  住院治疗;  $30 \sim 60$  住院和家庭治疗;  $> 60$  家庭治疗)。

2. 操作步骤

- (1) 按【分析→生存函数→Cox 回归】顺序单击菜单项, 打开如图 18-12 所示对话框。
- (2) 从左侧的变量表中选择 time 变量, 送入右侧的【时间】框中。
- (3) 选择 status 变量送入【状态】框中。单击【定义事件】按钮, 在打开的对话框【单值】框中输入“0”。
- (4) 选择 age、cell、diagtime、kps、prior、therapy 作为协变量送入【协变量】框。
- (5) 在【方法】下拉列表中选择【向后: Wald】项。
- (6) 单击【分类】按钮, 打开相应的对话框。选择 cell、prior、therapy 变量进入【分类协变量】框中。选中这三个变量, 使它们的【对比】方式均为【指示符】, 其中 cell 变量的【参考类别】为【第一个】, 其他两个分类变量为【最后一个】。
- (7) 单击【选项】按钮, 展开相应的对话框, 见图 18-16。选中【估计值的相关性】, 在【显示模型信息】栏内选择【在最后一个步骤中】, 在【进入】和【删除】框内分别输入“0.05”和“0.10”, 在【最大迭代数次数】框中输入“20”。
- (8) 单击【确定】按钮, 提交计算。

3. 输出结果

表 18-10 所示是对数据处理说明。不含删失数据的观测为 128, 含有删失数据的观测数为 9, 带有负生存时间的观测数 0, 在分层中删失观测数 0, 去除的观测总数 0, 用于统计分析的观测数 137, 以及它们占总观测数的百分比。

表 18-11 所示是各变量值编码。cell 分类变量, 以该变量中的第一分类(即 squamous)作为参照分类(编码 0、0、0)。(1)代表 small 类, (2)代表 adeno 类, (3)代表 large 类。

表 18-12 所示为第一步全模型与最后一步模型对系数检验的对数似然比值、总体分数的卡方检验、从前一步到本步变化量的卡方检验等。

表 18-13 中使用向后剔除拟合的第一步和最后一步的统计量和沃德检验, 步骤 1 全部指定的协变量进入模型, 但 Wald 检验说明只有 cell、kps 两变量对模型贡献显著步骤 5 说明经过一步步剔除对模型没有统计意义的协变量, 最后剩下 cell、kps。

表 18-10 观测处理表

案例处理摘要		N	百分比
分析中可用的案例	事件 <sup>a</sup>	128	93.4%
	删失	9	6.6%
	合计	137	100.0%
删除的案例	带有缺失值的案例	0	0.0%
	带有负时间的案例	0	0.0%
	层中的最早事件之前删失的案例	0	0.0%
	合计	0	0.0%
	合计	137	100.0%

a. 因变量: 生存时间

表 18-11 各变量值编码

分类变量编码 <sup>a,c,d</sup>		频率	(1) <sup>e</sup>	(2)	(3)
therapy <sup>b</sup>	1=标准方法	69	1		
	2=实验方法	68	0		
cell <sup>b</sup>	1=鳞癌	35	0	0	0
	2=小细胞肺癌	48	1	0	0
	3=腺癌	27	0	1	0
	4=大细胞肺癌	27	0	0	1
prior <sup>b</sup>	0=经过处理	97	1		
	1=未经处理	40	0		

- a. 分类变量: therapy (治疗方案)
- b. 示性参数编码
- c. 分类变量: cell (肺癌细胞组织学分类)
- d. 分类变量: prior (治疗前处理)
- e. 已经记录了 (0,1) 变量, 所以其系数不会与指示符 (0,1) 编码相同。

表 18-12 模型系数综合检验

模型系数的综合测试 <sup>b</sup>										
步骤	-2 倍对数似然值	整体(得分)			从上一步骤开始更改			从上一块开始更改		
		卡方	df	Sig.	卡方	df	Sig.	卡方	df	Sig.
1 <sup>a</sup>	950.359	65.917	8	.000	61.409	8	.000	61.409	8	.000
5	952.997	63.219	4	.000				58.771	4	.000

a. 在步骤编号 1: age cell kps diagtime prior therapy 处输入变量

b. 起始块编号 1. 方法 = 向后逐步 (Wald)

表 18-13 进入方程变量的统计量

方程中的变量						
	B	SE	Wald	df	Sig.	Exp(B)
步骤 1						
age	-.009	.009	.844	1	.358	.991
cell			17.916	3	.000	
cell(1)	.856	.275	9.697	1	.002	2.355
cell(2)	1.188	.301	15.610	1	.000	3.281
cell(3)	.400	.283	1.999	1	.157	1.491
kps	-.033	.006	35.112	1	.000	.968
diagtime	.000	.009	.000	1	.992	1.000
prior	-.072	.232	.097	1	.755	.930
therapy	-.290	.207	1.958	1	.162	.748
步骤 5						
cell			17.080	3	.001	
cell(1)	.712	.253	7.939	1	.005	2.038
cell(2)	1.151	.293	15.441	1	.000	3.161
cell(3)	.325	.277	1.381	1	.240	1.384
kps	-.031	.005	35.612	1	.000	.970

表 18-14 未进入模型变量的统计量

不在方程中的变量 <sup>a</sup>			
步骤 5	得分	df	Sig.
age	.424	1	.515
diagtime	.165	1	.684
prior	.248	1	.618
therapy	1.650	1	.199

a. 残差卡方 = 带有 4 df Sig. 的 2.675。 = .614

表 18-16 协变量均值

协变量均值	
	均值
age	58.307
cell(1)	.350
cell(2)	.197
cell(3)	.197
kps	58.569
diagtime	8.774
prior	.708
therapy	.504

表 18-15 回归系数相关矩阵

回归系数的相关矩阵			
	cell(1)	cell(2)	cell(3)
cell(2)	.570		
cell(3)	.517	.471	
kps	.159	.014	-.097

表 18-14 所示为拟合结束时，未进入模型变量的统计量。检验结果 Sig.都大于 0.05，表明对模型无统计意义的变量都没有进入模型。

表 18-15 所示为回归系数的相关矩阵。相关系数均不大，说明进入模型的变量之间相互基本是独立的，共线性问题不明显。

表 18-16 所示为协变量均值。

以上统计结果表明，kps 和 cell 变量对模型有显著性意义。kps 变量相对危险度为 0.970，回归系数为-0.031，说明 kps 变量取值越大，生存时间越长。在 cell 变量中，adeno 和 small 分类与 squamous 分类相比具有显著性，而 large 与 squamous 相比不具有显著性差异。adeno 的回归系数为 1.151，相对危险度为 3.161；small 回归系数为 0.712，相对危险度为 2.038；large 的回归系数为 0.325，相对危险度为 1.384。所以鳞癌细胞肺癌患者生存时间最长，其次是大细胞肺癌患者，再次是小细胞肺癌患者，腺癌细胞肺癌患者的生存时间最短。

18.5 Cox 依时协变量回归模型分析

在预后因素对其死亡风险的作用强度在所有时间上不能都保持一致时，也就是说，风险比率随时间而变化，在不同的时间点上一个(或多个)协变量有不同的值时，就需要使用扩展的“Cox 回归”模型，也就是 Cox 依时(更多地将其称为时间依存)协变量回归模型。在该模型中，需先指定时间依存协变量，再将其作为协变量作 Cox 回归分析。

18.5.1 Cox 依时协变量回归分析过程

1. 计算时间依存变量

按【分析→生存函数→Cox 依时协变量】顺序单击菜单项，打开如图 18-18 所示【计算时间依存变量】对话框。

在此对话框中,可在【T\_COV\_的表达式】框中设定逻辑表达式,以此来计算时间依存变量。

在左侧的框中,列出了当前数据文件中的所有变量,这些变量可供用于构造时间依存变量。其中一个名为 T\_的变量是系统另外提供的时间变量。它可以与其他变量结合起来构建时间依存变量。在构建时间依存变量表达式的过程中,可以充分使用右侧的各个键和 SPSS 提供的各种函数。

时间依存变量的构建取决于以下两种情形:

(1) 如果要检验关于特定的协变量的比例风险假设或者估计一个非比例风险的扩展“Cox 回归”模型,则可以将时间依存变量定义为该协变量和时间变量 T\_的函数。常用的方法是把时间变量 T\_和该协变量简单相乘即可,当然还可以指定较为复杂的函数。通过对时间依存协变量系数的显著性检验就可以知道成比例的风险假设是否合理。

(2) 有些变量尽管在不同的时间点取不同的值,但其值与时间并非系统地相关。在这种情况下,需用逻辑表达式定义一个分段时间依存协变量。逻辑表达式为真时取值“1”,为假时取值“0”。用一系列的逻辑表达式,就可以从一系列观察记录中建立起自己的时间依存变量。例如,对住院 3 周的患者,每周测量 1 次血压,共观察 3 次(变量名为 BP1~BP3),则时间依存协变量可这样定义:

$$\text{Var} = (T_ < 1) * \text{BP1} + (T_ \geq 1 \& T_ < 2) * \text{BP2} + (T_ \geq 2 \& T_ < 3) * \text{BP3}$$

式中,&表示“逻辑与”。该函数意味着当时间不足 1 周时,此时第一个括号内取值为 1,其他括号内的取值为 0,故使用 BP1 的值;1~2 周时,中间括号内取值为 1,其他括号内的取值为 0,故使用 BP2 的值;显然,2~3 周时,使用 BP3 的值。

## 2. 模型设定

单击【模型】按钮,打开如图 18-18 所示的【Cox 回归:模型】主对话框,它与图 18-12 的【Cox 回归:风险比例模型】主对话框相比,除没有绘图功能及 Bootstrap 功能外,其余完全一样。因此,这里不再赘述,相关内容可参见 18.4 节。

### 18.5.2 Cox 依时协变量回归分析实例

【例 4】数据文件 data18-04.sav 来源于 SPSS 自带的文件 recidivism.sav 中的数据,它记录的是第二次被逮捕罪犯时罪犯年龄和相距第一次被逮捕的时间,以及包括其他信息的一个随机样本。

涉及分析的主要变量有:Time,标签为“Time to second arrest”,第二次被逮捕距第一次被逮捕的时间(天),数值型尺度变量;arrest2,标签为“Second arrest”,第二次被逮捕,值标签 0 表示 no,1 表示 yes,分类名义变量;age,标签为“Age in years”,年龄,数值型尺度变量。

政府执法机构非常关心其辖区的累犯率。测度累犯的指标之一是罪犯第一次逮捕到第二次逮捕之间的时间。该机构想使用 Cox 回归模型构建重新被逮捕时间的模型,以此来研究罪



图 18-18 【计算时间依存变量】对话框

犯在两次犯罪时间上的一些规律，但担心在整个年龄组别中年龄的比例风险不随时间变化的假定无效。现使用 Cox 回归评估关于年龄的比例风险假定。

操作步骤如下：  
按【分析→生存函数→Cox 回归】顺序单击菜单项，打开如图 18-12 所示对话框。  
在左侧变量名框中，选择 Time to second arrest 变量作为时间变量，将其移入【时间】框。  
选择 Second arrest 作为状态变量，将其移入【状态】框。  
单击【定义事件】按钮，在打开的【定义状态变量发生事件】对话框中，选择【单值】，并在其后框中输入值“1”，表示第二次监禁。  
选择 Age in years 作为协变量，将其移入【协变量】框。  
单击【继续】按钮，返回【Cox 回归】主对话框。  
单击【保存】按钮，在弹出的对话框中选择【偏残差】选项。单击【继续】按钮，返回【Cox 回归】主对话框。

单击【确定】按钮，在当前工作的数据文件中生成一系列用 PR1\_1 作为变量名、保存 Age in years 的偏残差。将该变量的度量标准(测度水平)修改为度量(尺度)。  
现在建立时间偏残差的散点图来检查比例风险的假定。

为建立时间偏残差的散点图，按【图形→图表构建程序】顺序单击菜单项，打开如图 18-19 所示的【图表构建程序】对话框。



图 18-19 【图表构建程序】对话框

在【库】的【选择范围】框中选择【散点图/点图】，并在右侧的图案库中选择简单散点图。单击第一张图案，并将其拖拽到【图表预览使用实例数据】框中。

在【变量】列表框中，选择变量部分残差 AGE[PR1\_1] 作为 y 轴变量，将其拖拽到【图表预览使用实例数据】框中的“是否为 y 轴？”虚框中；选择变量 Time to second arrest 作为 x 轴变量，将其拖拽到【图表预览使用实例数据】框中的“是否为 x 轴？”虚框中。单击【确定】按钮，则在输出窗中得到有关上述两变量的散点图，见图 18-20。

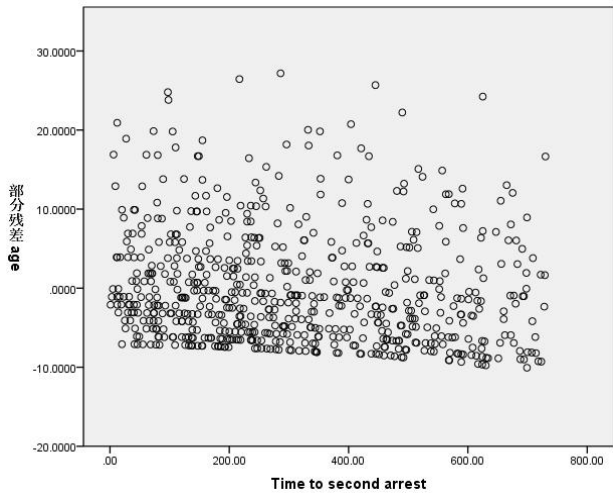



图 18-20 Time 与 PR1\_1 的散点图

为能看到重叠在散点图上的最佳拟合线，在输出窗中双击该散点图，打开使它处在激活状态的图表编辑器，见图 18-21。在图表编辑器中单击一个点，单击“添加总计拟合线”图标，则产生添加了拟合线的如图 18-21 所示的散点图，它可用来评估比例风险模型的假定。

水平轴显示直到第二次监禁的时间。垂直轴显示 Age in years 的偏残差。协变量的偏残差是每个个案协变量值的观察值和(给定模型为正确时)期望值之间的差异。

偏残差以及在图中这样的点仅仅是为无约束的个案产生的。  
如果关于罪犯的年龄的风险比例假定是正确的，则在本图中应不会有趋势性形态。可是，在偏残差和时间之间有一个清晰的负相关，这暗示 Age in years 的效应取决于时间。

如果遇到这种情况，则向模型中增加时间依存协变量。  
为运行含有时间依存协变量的 Cox 回归分析，按【分析→生存函数→Cox 依时协变量】顺序单击菜单项，打开如图 18-18 所示的【计算时间依存变量】对话框。

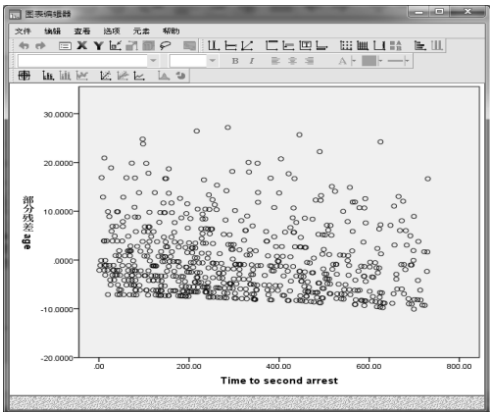


图 18-21 处在编辑状态的散点图

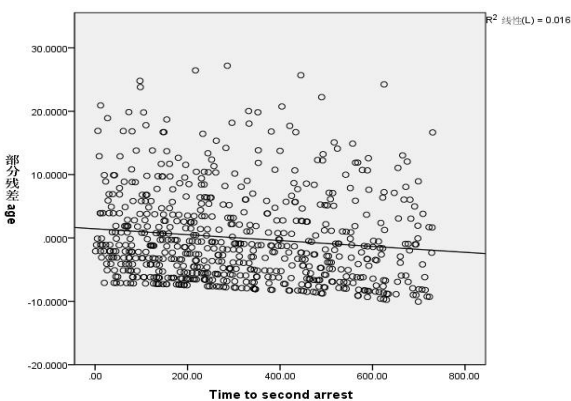


图 18-22 添加拟合线的散点图

在【T\_COV\_的表达式】框中输入时间依存协变量的表达式“T\_\*age”。  
单击【模型】按钮，打开如图 18-12 所示的【Cox 回归】主对话框。选择左侧变量名框中的 Time to arrest (time)，将其移入【时间】框，作为时间变量；选择 Second arrest (arrest2)，

将其移入【状态】框，作为状态变量。单击【定义事件】按钮，在展开的【定义状态变量发生事件】对话框中选择【单值】，并在其后框中输入值“1”，单击【继续】按钮，返回【Cox 回归】主对话框；选择 Select Age in years (age)和 T\_COV\_，将其移入到【协变量】框中，单击【确定】按钮，在输出窗中得到如表 18-17 所示的输出结果。

从表 18-17 中可见，时间依存协变量的显著性值小于 0.05，意味着它对模型有贡献，但系数数值很小。

为改变时间依存协变量系数显示的小数点位数，用双击表格的方法激活该表格。选择包含该系数的单元格，按【格式→单元格属性】顺序打开如图 18-23 所示的【单元格属性】对话框。

表 18-17 方程中的变量

方程中的变量						
	B	SE	Wald	df	Sig.	Exp(B)
age	-.026	.010	6.346	1	.012	.975
T_COV_	.000	.000	10.736	1	.001	1.000



图 18-23 【单元格属性】对话框

在【小数】框中输入“8”作为显示的小数位数。单击【确定】按钮，显示的系数值为 0.00009300。系数如此小的原因是时间依存协变量的值可以非常大。例如，对于一个罪犯 45 岁在第一次逮捕释放后 200 天的 T\_COV\_的值为  $45 \times 200 = 9000$ 。

为了使得该值不太极端，可以重新缩放时间轴为周、月或年。使用 Cox 回归程序，可以发现累犯年龄的效应是时间依存的，并向模型增加一项，有助于解释时间依存性。

习 题 18

1. 什么是寿命表和 Cox 模型？
2. 数据文件 data18-05 为 3 期和 4 期黑瘤患者的数据。其中，id 变量为编号，age 变量为年龄，sex 变量为性别(1: 男；2: 女)，survtime 变量为生存时间，survstatus 变量为生存状态(0: 死亡；1: 删失)，stage 变量为肿瘤级别。计算时间间隔为 5 个月的不同肿瘤级别寿命表。(数据来源：《生存数据分析的统计方法》，Elisa Lee 著，陈家鼎等译，北京：中国统计出版社，1998.)

3. 数据文件 data18-06 收集 63 例患者的生存时间、结局及影响因素。各变量的含义见表 18-18。请用 Cox 模型进行预后分析。(数据来源:《医学统计学》,孙振球主编,北京:人民卫生出版社,2002.)

表 18-18 某恶性肿瘤的影响因素及量化值

变 量	意 义	值标签(或单位)
X0	编号	
X1	年龄	岁
X2	性别	1, 男, 2, 女
X3	组织学类型	0, 低分化, 1, 高分化
X4	治疗方式	0, 新方法, 1, 传统方法
X5	淋巴结是否转移	0, 否, 1, 是
X6	肿瘤的浸润程度	0, 未突破浆膜, 1, 突破浆膜
t	生存时间	月
Y	患者结局	0, 死亡, 1, 截尾

# 第 19 章 生成统计图形

## 19.1 概 述

统计图是用点的位置、线段的升降、直条的长短或面积的大小等方法表达统计资料的一种形式，其特点是简明生动、形象具体和通俗易懂。

SPSS 制图功能很强，能绘制许多种统计图形，这些图形可以由各种统计分析过程产生，也可以直接从【图形】菜单中的一系列图形选项直接产生一部分。【图形】菜单提供三种图形产生方式。本章主要介绍通过旧对话框生成的统计图形。

SPSS 系统直接从当前数据窗口中读取指定数据而生成图形，数据影响图形的生成，因而在生成图形之前需要完成以下几个步骤。

### 1. 建立数据文件

打开数据窗口，录入有关数据。现有我国 12 座大中城市 1985—1994 年每月平均气温的数据文件，数据文件结构是以各个城市的月平均气温、年代和月份作变量。本资料来源于 1986—1995 年《中国统计年鉴》（中国统计出版社），数据文件 data19-01。

### 2. 制定数据文件结构

数据文件结构往往决定着图形的类型，同样是来源于同一个资料，可以做成不同的数据文件结构。例如，数据文件 data19-01 结构可以生成 1985—1994 年某个城市十二个月份平均气温图，但不能生成 1985—1994 年某个月份各城市平均气温图，必须改变数据文件 data19-01 的结构，以每个月份的平均气温、年代和城市作变量重新制作数据文件结构。数据文件编号 data19-02。

### 3. 调整数据文件结构

为了制图需要对已有的数据文件结构进行一些调整，就数据文件 data19-02 而言，若生成 1985—1994 年上海 4 月、5 月和 6 月的平均气温图，就必须将上海的数据单独生成一些变量。将凡是上海的数据复制，然后在当前的数据文件中粘贴数据形成有关上海气温的变量，也可以在一个新数据文件中粘贴这些数据形成有关的变量，数据文件编号为 data19-03。为了区别已有的变量，将有关上海变量的变量名加上“sh”。

## 19.2 条形图和 3D 条形图

条形图（Bar Charts）是利用相同宽度条形的长短或高低表现统计数据大小或变动的统计图，条形图还有其他别名，横排称为带形图，纵排又称柱形图。平面条形图只能显示两个变量，而 3D 条形图可以同时显示三个变量。它常用于分类资料的图示。



## 19.2.1 选择图形类型

用户在条形图生成时，首先选择图形类型。

按【图形→旧对话框→条形图】顺序单击菜单项，打开【条形图】对话框，见图 19-1。

按【图形→旧对话框→3D 条形图】顺序单击鼠标，打开【3D 条形图】对话框，见图 19-2。



图 19-1 【条形图】对话框

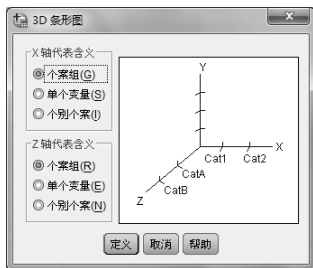


图 19-2 【3D 条形图】对话框

### 1. 条形图图式

- (1) 【简单箱图】。以若干平行且等宽的矩形表现数量对比关系，条间有间隙。
- (2) 【复式条形图】。由两个或两个以上条组成一组的条形图。
- (3) 【堆栈面积图】。它是以条形的全长代表某个变量的整体，条内的各分段长短代表各组成分在整体中所占的比例，每一段用不同线条或颜色表示。

### 2. 图表中的数据

- (1) 【个案组摘要】。每组观测量生成一个简单、分类、分段图形。
- (2) 【各个变量的摘要】。这种模式至少要选择两个或两个以上相同或不同的变量。
- (3) 【个案值】。每个观测值生成一个图形。

在【3D 条形图】对话框中，X 轴和 Z 轴代表含义分别为【个案组】、【单个变量】和【个别个案】，其功能与上述三个选项相对应。

## 19.2.2 简单条形图

这是一种对分类变量进行说明的条形图，条的高度代表分类变量本身情况。

读取数据文件 data19-04，在【条形图】对话框中选择【简单箱图】和【个案组摘要】，单击【定义】按钮，打开【定义简单条形图】对话框，见图 19-3。

在该对话框中定义图形参数。

- (1) 【类别轴】框。设置分类轴变量。

在变量列表框中选择 cont 为分类轴变量，送入该框。默认的分类轴是横轴。

分类轴上各变量值的排列位置，是由分类变量中变量值的大小和字母顺序所确定的，数值最小或字母顺序最靠前的变量值排在分类轴的最左端，相反则排在最右端。

在变量列表框中选择纵轴变量移入【变量】框，本例选择“wine”。

- (2) 【条的表征】栏。选择条图表达的统计量，共两大类：一类是对分类变量的描述；另一类是对其他变量的描述。

① 分类变量的计数函数，表达某一变量值，包括【个案数】、【个案数的%(百分比)】、【累积个数】和【累积%(百分比)】。

② 【其他统计量】。

【变量】框中所显示的是统计函数表达式“Mean[wine]”，Mean 为统计函数，wine 为统计函数的自变量，默认条长表示葡萄酒产量均值。倘若选择其他统计函数，单击【更改统计量】按钮，打开如图 19-4 所示的【统计量】对话框，在对话框内选择【变量】框中表达式的统计函数部分。统计函数共 4 组 18 个选项。

第一组统计函数包括【值的均值】、【值的中位数】、【值的众数】、【个案数】、【值的和】、【标准差】、【方差】、【最小值】、【最大值】和【累计求和】。



图 19-3 【定义简单条形图】对话框



图 19-4 【统计量】对话框

第二组统计函数与【值】框中的参数有关联，在【值】框中指定一个参数后，以下的函数为：

- 【上百分比】。大于指定参数的观测量数目占总数的百分比。
- 【下百分比】。小于指定参数的观测量数目占变量值总数的百分比。
- 【上个数】。大于指定参数的观测量数。
- 【下个数】。小于指定参数的观测量数。
- 【百分位】。指定参数的百分位数。

以 cont 变量作为分类变量，在【统计量】对话框中选择【其他统计量】，将 wine 变量选入【变量】框，单击【更改统计量】按钮，在【统计量】对话框中选择【上百分比】项，并在【值】框中输入“100”，生成 1988—1992 年各洲葡萄酒产量大于 100 万升的国家占该地区国家总数的百分比条形图，见图 19-5(a)。

在【统计量】对话框中选择【下个数】项，并在【值】框中输入“100”，生成 1988—1992 年各洲葡萄酒产量小于 100 万升的国家数量对比条形图，见图 19-5(b)。

第三组统计函数包括两个函数选项。在【低】框和【高】框内分别指定下限、上限值，两个参数在自变量值的范围内，且低值小于高值。生成图形的分类轴包括低值和高值两个点。生成图形剔除了自变量中的缺失值。

- **【内百分比】**。落在**【低】**框和**【高】**框参数范围内的观测量数占总数的百分比。
- **【内个数】**。落在**【低】**框和**【高】**框参数范围内的观测量数目。

第四组仅有一个**【值是组中点】**选项，变量值以中点分组，若选择了**【值的中位数】**和**【百分位数】**，该选项有效。选中此项，计算中位数和百分位数。

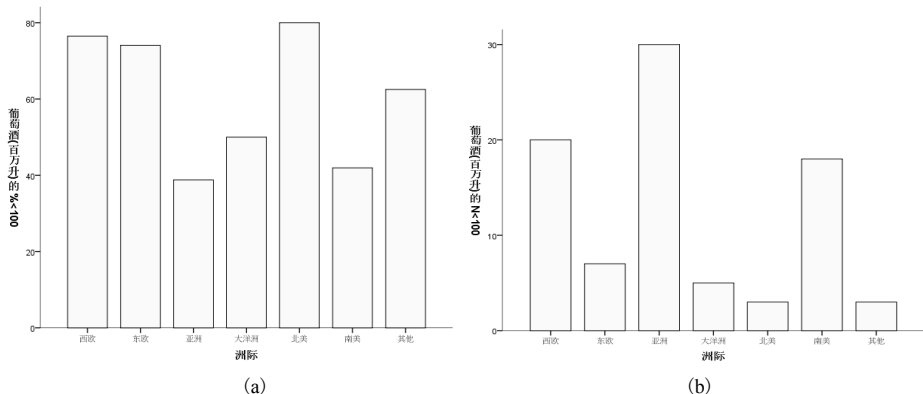


图 19-5 例图

(3) 单击**【标题】**按钮，出现**【标题】**对话框，见图 19-6 所示。具体操作参见**【图形编辑】**窗口的**【标题】**命令。

(4) 单击**【选项】**按钮，打开如图 19-7 所示的**【选项】**对话框。



图 19-6 【标题】对话框



图 19-7 【选项】对话框

① **【缺失值】**栏。选择缺失值处理方式。

- **【按列表排除个案】**。如果个案在一变量中有缺失值，那么剔除该个案。
- **【按变量顺序排除个案】**。某个变量中存在缺失值，仅剔除该变量的缺失值。
- **【显示缺失值所定义的组】**。在图形中显示缺失值所在的分组。

② **【使用个案标签显示图标】**。在图形中显示个案的标签值。

③ **【误差条图的表征】**栏。即误差条图所表达的统计量。

- **【置信区间】**。在**【级别】**框输入需要的水平值。
- **【标准误】**。在**【乘数】**框中根据需要输入标准误的倍数。
- **【标准差】**。在**【乘数】**框中根据需要输入标准差的倍数。

(5) 【图表规范的使用来源】。选择该项后，单击【文件】按钮，出现【从文件模板】对话框，指定模板文件。新生成的图形其大小、比例、小数位数、字形、字体以及图题的位置等都自动转换成模板格式。通过套用已经做好的图形模板，使生成的图形风格一致。

19.2.3 复式条形图

读取数据文件 data19-05。在【条形图】对话框中选择【复式条形图】和【各个变量的摘要】，单击【确定】按钮，打开相应对话框，见图 19-8。生成不同年龄不同性别血压变化图，见图 19-9。

(1) 【条的表征】框。在条形图表达统计量框中至少要有两个或两个以上的变量。变量在本框中上下的位置，决定着这些被选变量在分类轴上从左向右排列的顺序。本小节选择收缩压(sp)和舒张压(dp)变量进入【条的表征】框。

(2) 【面板依据】栏。生成群组图形，由若干按照一定方式排列的小图形组成。

① 【行】框。确定横向排列图形的变量，将性别(sex)变量选入该框。

② 【列】框。确定纵向排列图形的变量，将年龄(age)变量选入该框。

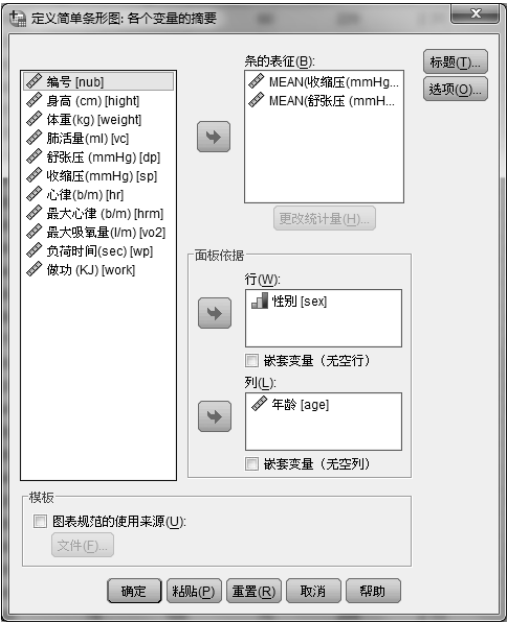


图 19-8 【定义简单条形图：各个变量的摘要】对话框

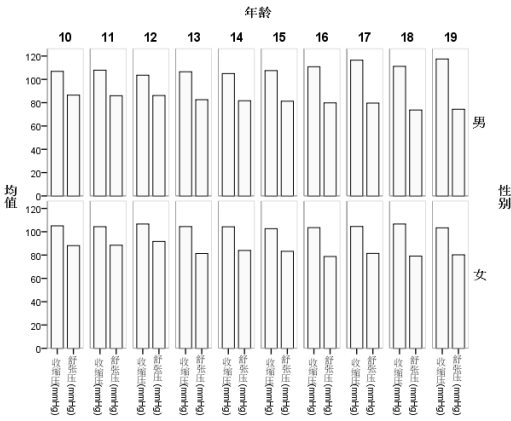


图 19-9 不同年龄不同性别血压变化图

19.2.4 堆积面堆图

(1) 读取数据文件 data19-04。

(2) 在【条形图】对话框中选择【堆积面积图】和【个案组摘要】，单击【确定】按钮，打开堆栈条形图对话框，见图 19-10。

(3) 【条的表征】栏中选择【其他统计量】项，选 cc 变量送入变量栏。

(4) 选择 cont 作为分类轴变量送入【类别轴】框。

(5) 选择 year 作为堆栈变量，送入【定义堆栈】框。堆栈是以堆栈变量中各变量值的数字或字母顺序排列，数值小或字母顺序靠前的变量值在条形图的下端，反之在条形图的上端。

单击【确定】按钮，生成 1988—1992 年各洲每年碳酸饮料平均产量，见图 19-11。



图 19-10 【定义堆积条形图：个案组摘要】堆栈条形图对话框

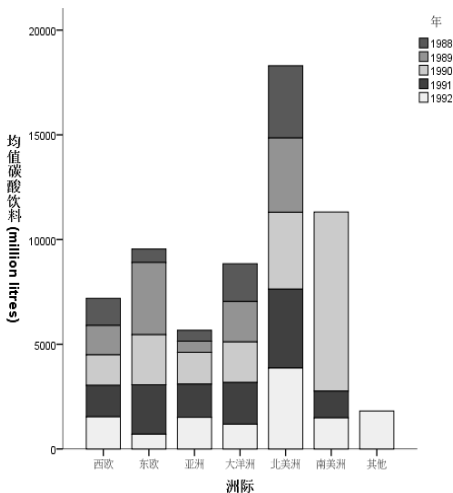


图 19-11 1988—1992 年各洲每年碳酸盐和浓缩饮料平均产量

## 19.2.5 3D 条形图

(1) 读取数据文件 data19-06。按【图形→旧对话框→3D 条形图】顺序单击菜单项，打开【定义 3D 条形图：个案组摘要】对话框，见图 19-12。

(2) SPSS 系统自动默认 Y 轴为数值变量轴，X 轴和 Z 轴分别为分类变量轴。在 X 和 Z 轴上分别选择【个案组】。

(3) 单击【确定】按钮，打开【定义 3D 条形图：个案组摘要】对话框，见图 19-12。将变量 salary 送入【变量】框中为 Y 轴变量。在【条的表征】下拉列表中选择【值的均值】项。选择 college 作为 X 轴变量送入【X 类别轴】框，Z 轴选择 gender 变量。

单击【确定】按钮，生成不同性别不同时期毕业生的初始薪酬对比图，见图 19-13。



图 19-12 【定义 3D 条形图：个案组摘要】对话框

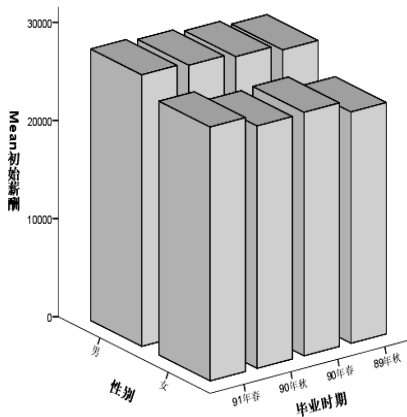


图 19-13 不同性别时期毕业生的初始薪酬

### 19.3 线图、面积图、高低图和圆图

线图又称曲线图，是用线段的升降来说明现象变动情况的一种统计图，它主要用于表示现象在时间上的变化趋势、现象的分配情况和两个现象之间的依存关系等。这里所指的线图均为纵横轴是算术刻度的普通线图。

面积图是用线段下的阴影面积来强调现象变化的统计图。堆栈面积图可表示现象总体内部结构状况，因此也称为结构曲线图。

高低图是一种说明某些现象在单位时间内变化情况的统计图。它适合描述每小时、每天、每周等时间内不断波动的市场信息资料，如股票、商品价格、货币牌价等，高低图既说明某些现象在短时间内的变化，也可说明它们长期的变化趋势。

饼图（Pie Charts）又称圆图，常用来表现构成比。以整个圆代表被研究现象的总体，按各构成部分占总体比重的大小把圆面积分割成若干扇形，表示部分对总体的比例关系。

#### 19.3.1 选择图形类型

按【图形→旧对话框→线图】顺序单击菜单项，打开【线图】对话框，见图 19-14。

按【图形→旧对话框→面积图】顺序单击菜单项，打开【面积图】对话框，见图 19-15。

按【图形→旧对话框→高-低图】顺序单击菜单项，打开【高-低图】对话框，见图 19-16。在各对话框中选择图形的模式。

按【图形→旧对话框→饼图】顺序单击菜单项，打开【饼图】对话框，见图 19-17。

(1) 在图 19-14 所示的【线图】对话框中选择线图模式。

- 【简单】。用一条折线表示某种现象变动趋势的统计图。
- 【多线线图】。用多条折线同时表示多种现象变动趋势的统计图。
- 【垂直线图】。反映某些现象在同一时期内差距的统计图。

(2) 在图 19-15 所示的【面积图】对话框中选择面积图模式。

- 【简单箱图】。用面积的变化表示某种现象变动趋势的统计图。
- 【堆积面积图】。用不同种类的面积表示多种现象变动趋势和总体内部构成。



图 19-14 【线图】对话框 图 19-15 【面积图】对话框 图 19-16 【高-低图】对话框 图 19-17 【饼图】对话框

(3) 在图 19-16 所示的【高-低图】对话框中选择高低图模式。

① 【简单高低关闭】。表示单位时间内某现象的最高数值、最低数值和收盘数值。它适用于股票、期货等，可说明每天最高、最低和收盘价。

②【简单范围栏】。表明单位时间内某现象最高数值和最低数值。它与单式高低收盘图的区别是省去了收盘值。

③【聚类高低关闭】。表示在单位时间内两个或两个以上现象的最高值、最低值和收盘值。

④【聚类范围栏】。表示在单位时间内两个或两个以上现象的最高值和最低值。

⑤【差别面积】。表示两个现象在同一时间内相互变化对比关系的线性统计图。

(4) 在图 19-17 所示的对话框中选择饼图模式。

统计量描述模式参见 19.2.1 小节。

### 19.3.2 堆积面积图

在【面积图】对话框中选择【堆积面积图】和【个案值】选项，单击【确定】按钮，打开【堆积面积图：个案的值】对话框，见图 19-18。本小节数据来自数据文件 data19-07，例图为 1950—1985 年我国每年国防支出总和和经济建设支出面积图，见图 19-19。主要操作步骤如下：

(1)【面积的表征】框。选 eco、def 变量送入此框，描述该变量。

(2)【类别标签】栏。

①【个案号】。以当前数据窗中的个案序号为标记排列【面积的表征】框内变量的变量值，分类轴上变量值用阿拉伯数字标记。

②【变量】。以某变量的变量值为标记排列【面积的表征】框内变量的变量值。在本框内选 year 变量。

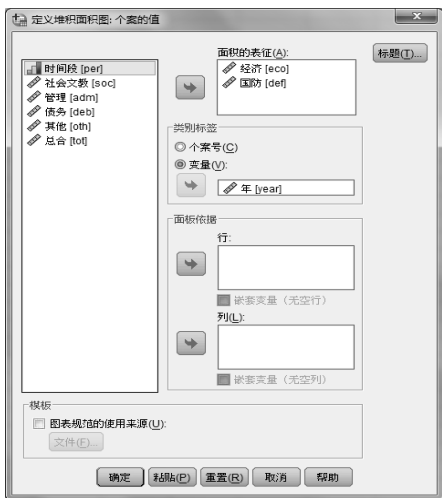


图 19-18 【定义堆积面积图：个案的值】对话框

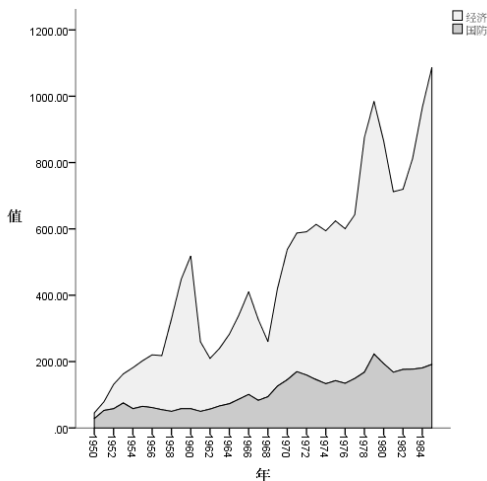


图 19-19 例图

### 19.3.3 多线线图

在【线图】对话框中选择【多线线图】和【个案组摘要】，单击【确定】按钮，打开【定义多线线图：个案组摘要】对话框，见图 19-20。使用数据文件 data19-01 中的数据，例图为 1985—1994 年武汉月平均气温变化图，见图 19-21。主要操作步骤如下：

(1)【线的表征】栏。选【其他统计量(例如均值)】项，选 wuhan 变量进【变量】框。

(2)【类别轴】框。选择 month 变量，具体操作见 19.2.2 小节。

(3)【定义线的方式】框。选择 year 变量。



图 19-20 【定义多线线图：个案组摘要】对话框

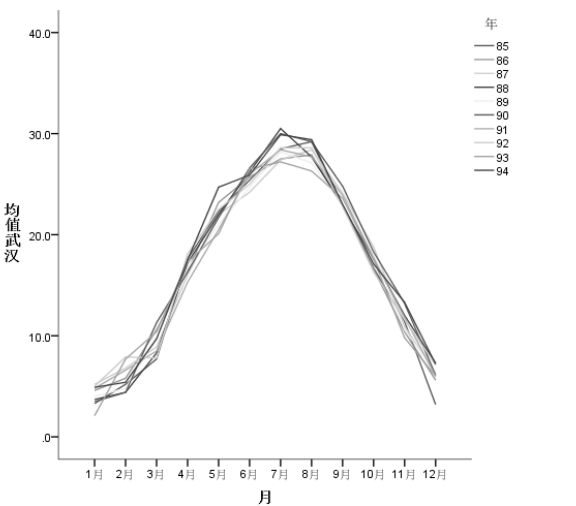


图 19-21 例图

19.3.4 垂线图

在【线图】对话框中选择【垂直线图】和【各个变量的摘要】项，单击【确定】按钮，展开【定义垂直线图：每个变量的摘要】对话框，见图 19-22。本小节数据来自数据文件 data19-01，例图为 1985—1994 年广州、北京、武汉月平均气温对比图，见图 19-23。主要操作步骤如下：

- (1) 将变量 beijing、guangzhou、wuhan 送入【点的表征】框中。
- (2) 将变量 month 作为分类轴变量送入【类别轴】框中，单击【确定】按钮。



图 19-22 【定义垂直线图：每个变量的摘要】对话框

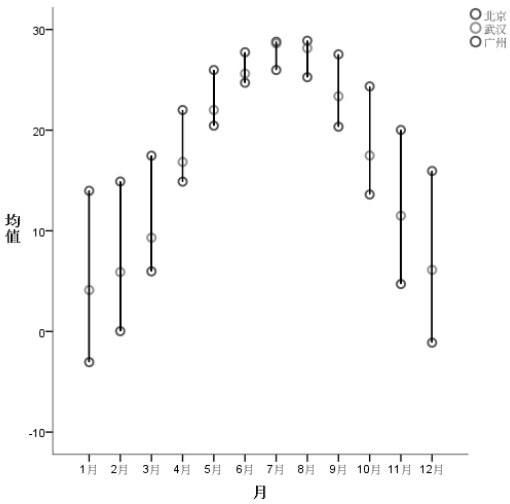


图 19-23 例图

19.3.5 简单高-低-闭合图

在【高-低图】对话框中选择【简单高低闭合】和【个案组摘要】选项，单击【确定】按钮，打开【定义简单高-低-闭合图：个案组摘要】对话框，见图 19-24。来自数据文件 data19-08，



例图为 1996 年 4 月 1 日至 19 日地产类股票每天最高价、最低价和收盘价变化图，见图 19-25。主要操作步骤如下：

(1) 【点的表征】栏。选择【其他统计量(例如均值)】项，再将 value 变量选入【变量】框，value 变量的统计量为 MEAN，改变统计函数参见 19.2.2 小节。



图 19-24 【定义简单高-低-闭合图：个案组摘要】

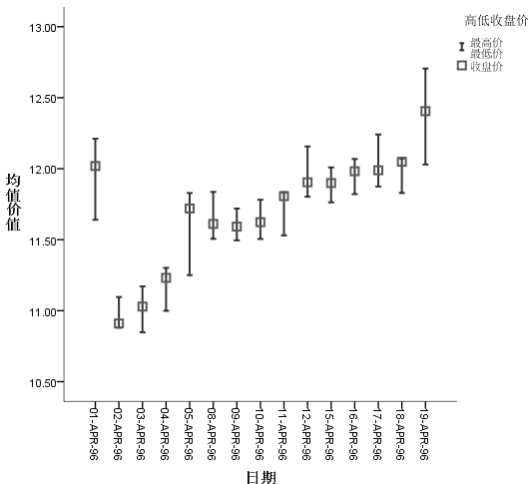


图 19-25 例图

(2) 【类别轴】框。选择 date 变量作为分类轴变量。

(3) 【定义高-低-闭合】框。选择 hlc 变量作为高低收盘变量，所生成条图的上端代表最高价，下端代表最低价，中间的方块代表收盘价。

### 19.3.6 聚类高低收盘图

在【高-低图】对话框中选择【聚类高低闭合(图)】和【各个变量的摘要】选项，单击【确定】按钮，打开【定义复式高-低-闭合图：各个变量的摘要】对话框，见图 19-26。本小节数据来自数据文件 data19-09，例图 19-27 为 1996 年第 14、15、16 周城乡股票、北人股票以及天桥股票对比变化图，见图 19-27。主要操作步骤如下：

- (1) 【高】框中的变量将作为条图的上端值。
- (2) 【低】框中的变量将作为条图的下端值。
- (3) 【闭合】框中的变量将作为条图的方块，是收盘价。
- (4) 【类别轴】框内的变量 week 作为分类轴。

【高】框和【低】框中必须选有变量，而【闭合】框则可选或不选入变量，如果在【闭合】框内没有选入变量，最后生成的图形就没有最后数值的标记（方块）。

【N 的变量集 M】，显示 N 套变量组中的第 M 套变量。当选择完一套变量后，即在【高】、【低】、【闭合】框中分别选入了一套变量的最高价、最低价或收盘价变量后，单击【下一张】按钮，出现提示录入下一套变量。本例中录入三套变量(chx-hi、chx-lo、chx-cl, tq-hi、tq-lo、tq-cl 和 br-hi、br-lo、br-cl)，录入完第一套 chx 变量，单击【下一张】按钮，出现“1 的变量集 1”提示；如果录入完这三套变量，文字提示显示“3 的变量集 3”，其含义为当前的这些变量是三套分组变量中

的第三套变量。要修改第二套变量，单击【上一张】按钮，文字提示显示“3 的变量集 2”，即为三套变量组中的第二套变量，同时在相应的变量框内显示第二套变量的高、低和关闭变量。

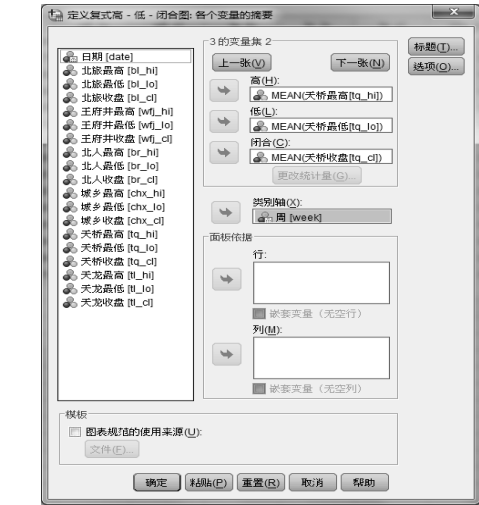


图 19-26 【定义复式高-低-闭合图：各个变量的摘要】对话框

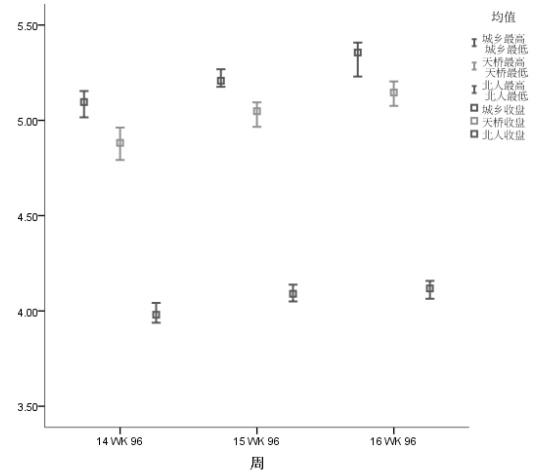


图 19-27 例图

19.3.7 简单极差图

在【高-低图】对话框中选择【简单范围（图）】和【个案组摘要】项，单击【确定】按钮，打开【简单极差图】对话框，见图 19-28。数据文件为 data19-10，例图为 1996 年 4 月 1 日至 19 日工业股票和商业股票每日收市平均价对比图，见图 19-29。主要操作步骤如下：



图 19-28 【简单极差图】对话框

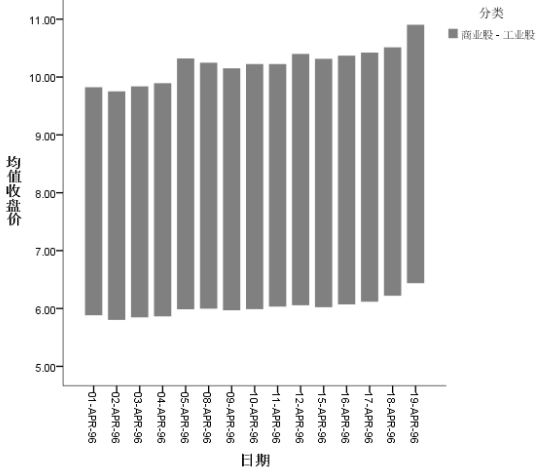


图 19-29 例图

- (1) 在【条的表征】栏中选择【其他统计量(例如均值)】项，将 close 变量选入【变量】框，close 变量的统计量为 MEAN。
- (2) 【类别轴】框内选入 date 作为分类轴变量。
- (3) 在【定义两个组】框中用分组变量来确定极差图两端的变量。选择 group 为分组变量。极差图两端各代表不同的变量值。本例中，从参与绘图的数据集中的原始数据来看，极差图下

端表示的是所选的工业股的每日均价，而极差图上端表示的是所选的商业股的每日均价，因此用来分组的变量只能有两个类别，通过极差图的长短表示工业股和商业股两个类别每日在 close 变量（收盘价）上均值之间的差距。

### 19.3.8 差分面积图

在【高-低图】对话框中选择【差别面积（图）】和【各个变量的摘要】选项，单击【确定】按钮，打开【定义差别面积图：各个变量的摘要】对话框，见图 19-30。例题数据文件 data19-01。例图为 1985—1994 年北京和天津年平均气温对比图，见图 19-31。图中的浅色代表天津年平均气温，深色代表北京年平均气温；浅色在上表示天津年平均气温高于北京年平均气温，而深色在上表示北京年平均气温高于天津年平均气温。

主要操作步骤是：将 Beijing 和 Tianjin 变量分别选入【第一个】和【第二个】框中，统计函数为 Mean；将 year 变量选入【类别轴】框中；单击【确定】按钮。



图 19-30 【定义差别面积图：各个变量的摘要】对话框

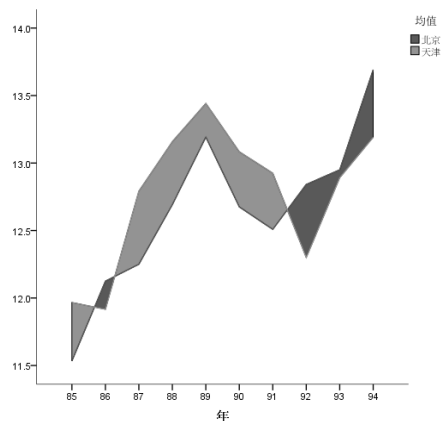


图 19-31 例图

### 19.3.9 饼图

在【饼图】对话框中选择【个案值】选项，单击【确定】按钮，打开【定义饼图：个案的值】对话框，见图 19-32。例题数据文件为 data19-11。例图为 1993 年乌克兰每月失业人口，见图 19-33。



图 19-32 【定义饼图：个案的值】对话框

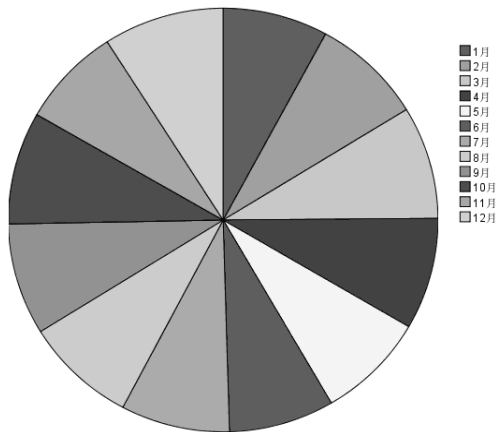


图 19-33 例图

## 19.4 箱图和误差条图

箱图 (Boxplots) 又称箱线图, 是一种描述数据分布的统计图形, 利用它可以从视觉的角度观察变量值的分布情况。箱图主要表示变量值的中位数、第 25 百分位数、第 75 百分位数等统计量, 其具体表示的统计量参见第 7 章探索分析一节。箱图可以从探索分析统计过程中获得, 但是本节介绍的方法能够制作更复杂的箱图。

误差条图 (Error Bar Charts) 是一种描述数据总体离散的统计图形, 利用它可以从视觉的角度观察样本的离散程度, 误差条图表达平均数的置信区间、标准差或标准误。在误差条图中, 小方块表示平均数, 图形的两端为置信区间、标准差或标准误。

### 19.4.1 选择箱图和误差条图类型

通过箱图对话框指定图形的类型, 按【图形→旧对话框→箱图】顺序单击菜单项, 打开【箱图】对话框, 见图 19-34。

通过【误差条图】对话框指定图形的类型, 按【图形→旧对话框→误差条图】顺序单击菜单项, 打开【误差条图】对话框, 见图 19-35。

箱图、误差条图图式和统计量描述模式, 请参看 19.2.1 节。根据箱图、误差条图图式和统计量描述模式的选择组合, 共可生成 4 种不同类型的箱图和误差条图。



图 19-34 【箱图】对话框



图 19-35 【误差条图】对话框

### 19.4.2 简单箱图

在【箱图】对话框中选择【简单箱图】和【个案组摘要】项, 单击【确定】按钮, 打开【定义简单箱图: 个案组摘要】对话框, 见图 19-36。例题数据文件为 data19-12。例图为不同岗位银行职员当前工资的箱线图, 见图 19-37。主要操作步骤是: 选择要描述的变量 salary 送入【变量】框; 选择分类轴变量 jobcat 送入【类别轴】框; 选择标识观测量的变量 gender 送入【标签个案依据】框。该变量值将对箱体外的观测量进行标识, “男”标识男性, “女”标识女性, 见图 19-37。

### 19.4.3 复式箱图

在【箱图】对话框中选择【复式条形图】和【各个变量的摘要】项, 单击【确定】按钮, 打开【定义复式箱图: 各个变量的摘要】对话框, 见图 19-38。数据文件 data19-12 中, 分类轴变量选择“职务等级”, 作箱图的变量选择“当前工资”、“初始工资”。例图为不同职务银行职

员初始工资和当前工资的箱线图，见图 19-38。如果没有选择标签变量，图中的数字为个案的编号。



图 19-36 【定义简单箱图：个案组摘要】对话框

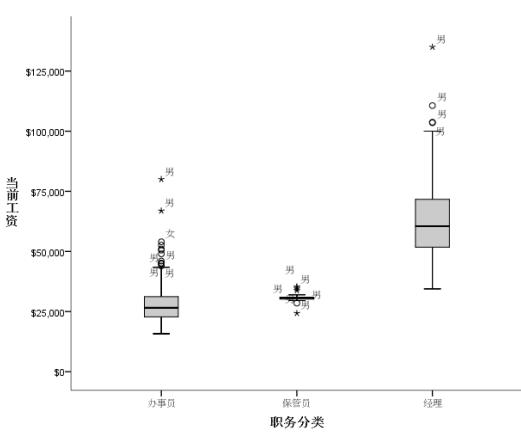


图 19-37 例图

主要操作步骤是：选择初始工资 salbegin 和当前工资 salary 两个变量作为箱图要描述的变量送入【框的表征】框中；选择雇员职务 jobcat 变量作为分类轴变量送入【类别轴】框中；单击【确定】按钮。

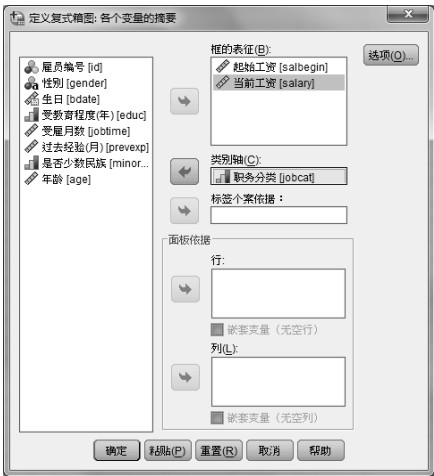


图 19-38 【定义复式箱图：各个变量的摘要】对话框

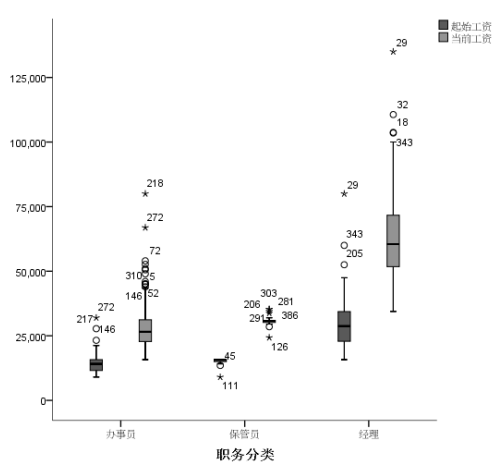


图 19-39 例图

### 19.4.4 简单误差条图

在【误差条图】对话框中选择【简单】和【个案组摘要】选项，单击【确定】按钮，打开【定义简单误差条形图：个案组摘要】对话框，见图 19-40。使用数据文件 data19-05 的数据。例图为各年龄组受试者体重均值 95%置信区间的误差条图，见图 19-41，中间的方块纵坐标是对应年龄的体重均值，上下横线是均值的 95%置信区间的上下限。主要操作步骤是：选择 weight 作为被描述变量送入【变量】框；选择 age 作为分类轴变量送入【类别轴】框。

在【条的表征】下拉列表中有 3 个选项：

(1)【均值置信区间】。在【度：】框中输入需要的置信区间。本例选择此项，在【度：】框中输入“95”。



图 19-40 【定义简单误差条形图：个案组摘要】对话框

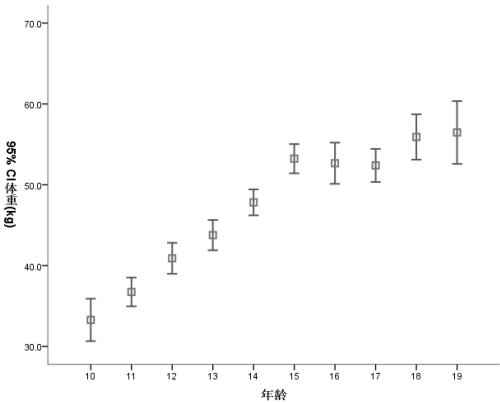


图 19-41 例图

- (2) 【均值标准误】。在【乘数】框中输入均值标准误的倍数。
- (3) 【标准差】。【乘数】框中可根据需要输入标准差的倍数。

19.4.5 复式误差条图

在【误差条图】对话框中选择【复式条形图】和【个案组摘要】项，单击【确定】按钮，打开【定义复式误差条形图：个案组摘要】对话框，见图 19-42。例题数据文件为 data19-05，例图为男女各年龄组身高两倍标准差范围的误差条图，见图 19-43。

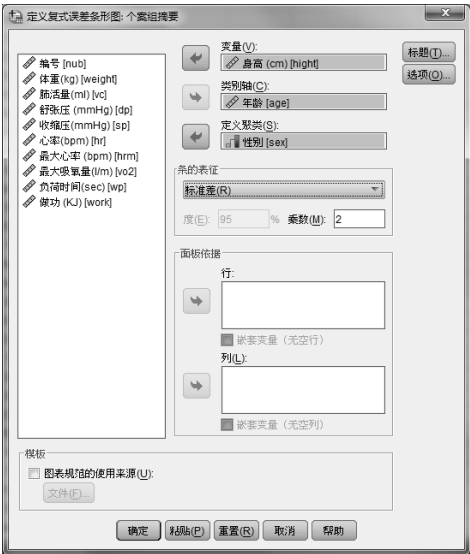


图 19-42 【定义复式误差条形图：个案组摘要】对话框

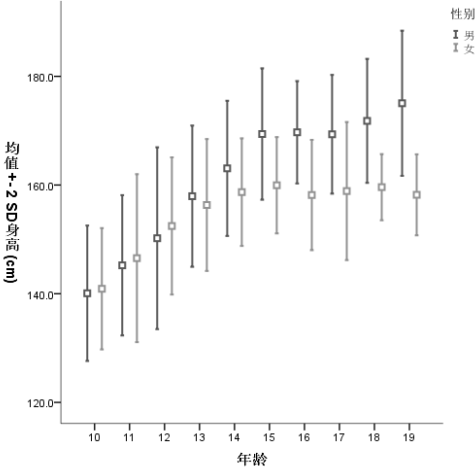


图 19-43 例图

主要操作步骤是：选择身高 height 变量送入【变量】框作为要描述的变量；选择年龄 age 变量作为分类轴变量送入【类别轴】框中；选择性别 sex 变量作为标识类别的变量送入【定义聚类】框中。在【条的表征】下拉列表中选择【均值标准误】，单击【确定】按钮。

## 19.5 散 点 图

散点图 (Scatterplots) 又称散布图或相关图, 是以点的分布反映变量间相关情况的图形, 根据图中的各点分布走向和密集程度, 大致可以判断变量之间协变关系的类型。

### 19.5.1 选择散点图图式

读者通过散点图对话框指定散点图图式, 按【图形→旧对话框→散点/点状】顺序单击菜单项, 打开【散点图/点图】对话框, 见图 19-44, 共有 5 种散点图。

- (1) 【简单分布】。显示一对相关变量的散点图。
- (2) 【重叠分布】。可显示多对相关变量的散点图。
- (3) 【矩阵分布】。在矩阵中显示多个相关变量之间的散点图。



图 19-44 【散点图/点图】对话框

- (4) 【3D 分布】。显示 3 个相关变量之间的散点图。
- (5) 【简单点】。每个点代表 1 个观测量, 在图形中显示

数值变量中各观测量在 X 轴上分布的图形, 该图也可看作一种散点图。

### 19.5.2 简单散点图

在【散点图/点图】对话框中选择【简单分布】项, 单击【定义】按钮, 打开【简单散点图】对话框, 见图 19-45。例题数据文件为 data19-05, 例图为男女受试者最大吸氧量与负荷时间的简单相关图, 见图 19-46。主要操作步骤是: 选择 wp 作为 Y 轴变量送入【Y 轴】框; 选择 vo2 作为 X 轴变量送入【X 轴】框; 选择 sex 送入【设置标记】框。

还可以选择标识个案的变量送入【标注个案】框。单击【选项】按钮, 打开如图 19-7 所示的【选项】对话框, 确定是否显示观测量的标识, 选中【使用个案标签显示图标】项, 标识变量才有效。



图 19-45 【简单散点图】对话框

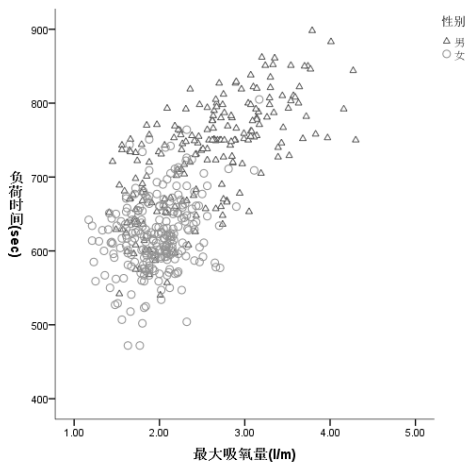



图 19-46 例图

19.5.3 重叠散点图

在【散点/点图】对话框中选择【重叠分布】选项，单击【定义】按钮，打开【重叠散点图】对话框，见图 19-47。例题数据文件为 data19-05，例图为舒张压与体重、做功量与体重、身高与体重的重叠相关图，见图 19-48。

在变量框中选择 Y-X 轴配对变量。第一个选择的为 Y 轴变量，第二个选择的为 X 轴变量。送入【Y-X 对】框。本例选择了 dp-weight（收缩压-体重）、work-weight（做功-体重）和 height-weight（身高-体重）3 个变量对。如果想要调换 Y-X 轴变量的位置，则先选择变量对，再单击  置换 Y-X 按钮。

一个图中表现了 3 对变量的关系。可以看出：身高-体重有明显的线性关系；做功与体重的线性关系不明显；而舒张压与体重几乎没有线性关系。



图 19-47 【重叠散点图】对话框

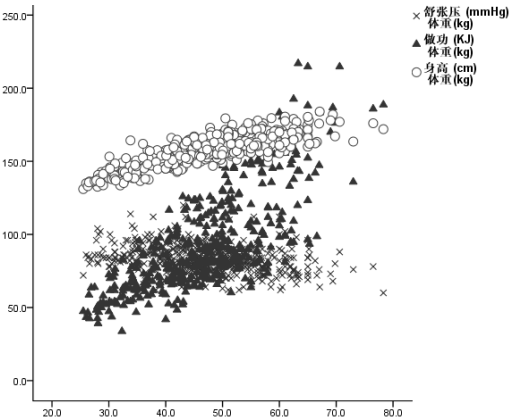


图 19-48 例图

19.5.4 矩阵散点图

在【散点图/点图】对话框中选择【矩阵分布】选项，单击【定义】按钮，打开【散点图矩阵】对话框，见图 19-49。例题数据文件为 data19-05。例图为男女受试者最大吸氧量、肺活量和最大心率矩阵散点图，见图 19-50。



图 19-49 【散点图矩阵】对话框

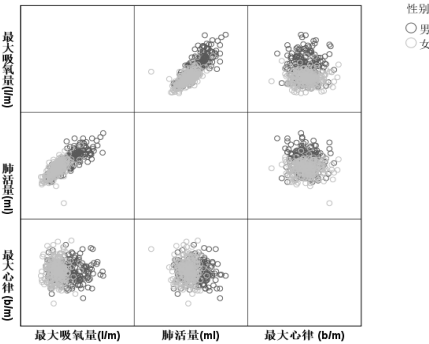


图 19-50 例图



(1) 【矩阵变量】框内要选择两个或两个以上的变量，本例选择 vo2、vc 和 hrm 变量作为被描述变量。请读者注意【矩阵变量】框内的变量顺序与矩阵散点图对角线变量的顺序。

(2) 【设置标记】框中设定散点标记，参见 19.8.2 小节。本例选择 sex 变量作为散点标记。  
从图 19-50 中可以看出，肺活量和最大吸氧量之间有明显较明显的线性关系；而最大心率与肺活量、最大吸氧量之间没有线性关系。

19.5.5 简单点图

在【散点图】对话框中选择【简单点】，单击【定义】按钮，打开【定义简单点图】对话框，见图 19-51。例题数据文件为 data19-05。

主要操作步骤是：选择 vo2 变量作为被观测的变量送入【X 轴变量】框，该变量必须为数值型变量。通过点图在 X 轴上的堆栈情况，观察观测量的分布状态；选择性别变量送到【行】框中。单击【选项】按钮，打开【定义简单点图：选项】对话框，见图 19-52，确定点图的分布形状：

- (1) 【不对称】。散点分布在 X 轴上侧。本题选择此项。
- (2) 【对称】。散点对称性地分布在 X 轴两侧。
- (3) 【水平】。散点平行地分布 X 轴上。

例图为男女受试者最大吸氧量在 X 轴上方的非对称性堆栈分布点图，见图 19-53。



图 19-51 【定义简单点图】对话框



图 19-52 【定义简单点图：选项】对话框

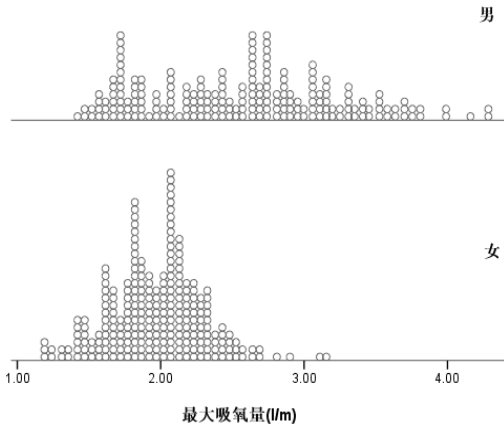


图 19-53 例图

19.6 直 方 图

直方图(Histogram)是以一组无间隔的直条,表现定量变量频数分布特征的统计图。直方图的每个条的高度代表相应组别的频数,可以直观地观察变量值的分布状况。



图 19-54 【直方图】对话框

本节数据文件为 data19-13,是某市 150 名 3 岁女童身高(cm),数据来源于《卫生统计》(周士楷,人民卫生出版社);数据文件 data19-14 是 1971 年某市调查 190 例正常人血铅含量( $\mu\text{g}/100\text{g}$ ),数据来源于《中国医学百科全书·医学统计学》(上海科学技术出版社)。

按【图形→旧对话框→直方图】顺序单击菜单项,打开【直方图】对话框,见图 19-54。

(1) 【变量】框。选择被描述的变量送入此栏。

① 使用数据文件 data19-13,选择变量 height 作描述变量。生成图 19-55(a),为带有正态曲线的某市 150 名 3 岁女童身高直方图。

② 使用数据文件 data19-14,选择变量 pb 做描述变量。生成图 19-55(b),为带有正态曲线的某市 190 例正常人血铅含量直方图。

(2) 【显示正态曲线】。在生成的直方图上显示正态曲线。图 19-55 中的两个例图都选择了此项。

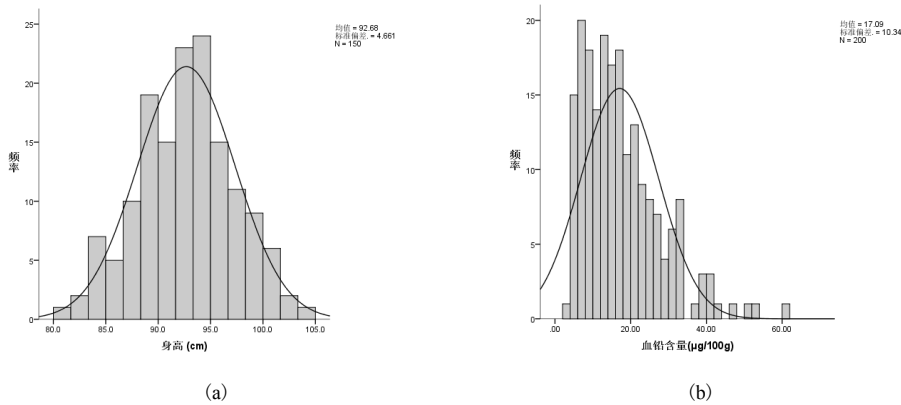


图 19-55 例图

可以看出,3 岁女童身高基本符合正态分布的特征;而血铅含量变量值呈非正态分布,有一个比较长的右尾,即血铅含量越高的人数越少,大部分人的血铅含量比较低。

19.7 帕 累 托 图

帕累托图(Pareto Charts)又称排列图或主次因素图,是作为改善质量管理活动中选择关键问题的一种工具。由于关键的多数和次要的多数现象具有普遍性,所以帕累托图也广泛应用于其他研究领域。

## 19.7.1 选择帕累托图类型

按【分析→质量控制→排列图】顺序单击菜单项，打开【帕累托图】对话框，见图 19-56。

### 1. 帕累托图图式

(1) 【简单】。它对分类轴上每一种类型的变量产生一个条形图，并按各种因素发生次数的多少从左到右顺序排列，帕累托曲线对分类轴上的每个变量值进行累加。

(2) 【堆积面积图】。是由分段条形图和帕累托图曲线构成的统计图。



图 19-56 【帕累托图】对话框

### 2. 统计量描述模式

(1) 【个案组的计数或和】。统计分类轴上的不同观测值数目，或对分类轴上观测值累加。

(2) 【单独变量的和】。累加分类轴上每个变量的值。

(3) 【个案值】。对分类轴变量中的每种观测值累加。

## 19.7.2 简单帕累托图

### 1. 个案组计数的简单帕累托图

在【帕累托图】对话框中选择【简单】和【个案组的计数或和】项，单击【定义】按钮，打开【定义简单排列图：个案组的计数或和】对话框，见图 19-57。

(1) 【条的表征】栏。表达两种不同类型的变量：

① 【计数】。只适用于字符型变量。

② 【变量和】。适用于数值型变量，被选定的变量在框中显示。

(2) 【类别轴】框。选择分类轴变量框。



图 19-57 【定义简单排列图：个案组的计数或和】对话框

(3) 【显示累积线】。系统默认为选定状态。选定此项，显示帕累托曲线（累积曲线）。

根据数据文件 data19-15，在【条的表征】框中选择 Counts 变量，选择 cat 不合格产品分类变量作为类别轴变量，生成切削刀质量帕累托图，见图 19-58(a)。通常把累计百分比分为三部分：0～80%表示主要因素（A 类），80%～90%表示次要因素（B 类），90%～100%表示一般因素（C 类）。可以看出，主要因素是短料和裂缝，其余为次要因素。

根据数据文件 data19-16，在【帕累托图】对话框中仍选择【简单】和【个案组计数或和】，在定义对话框中【条的表征】栏中选择【变量和】项，选择 nurse 数值型变量进入框中，并选择 cont 变量作为分类轴变量，生成的图形为各洲护士人数排列图，见图 19-58(b)。

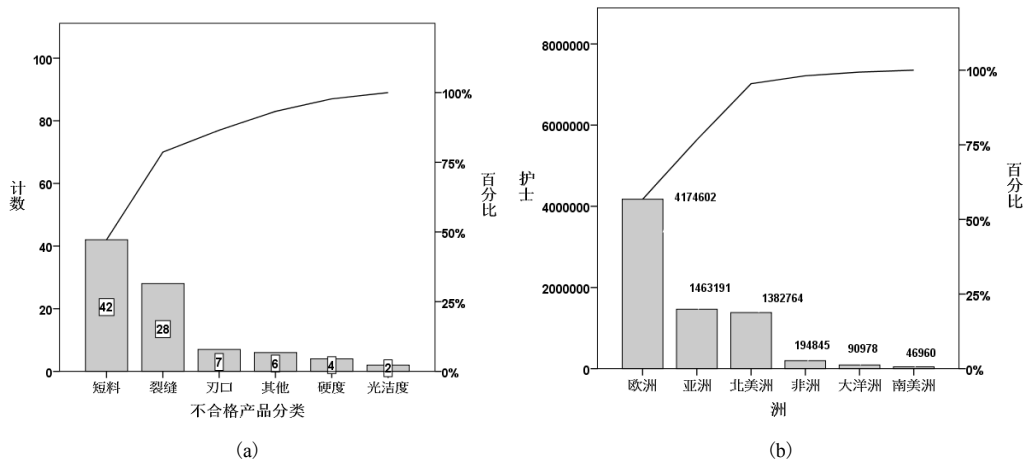


图 19-58 例图

2. 个案值的简单帕累托图

在【帕累托图】对话框中选择【简单】和【个案值】项，单击【定义】按钮，展开【定义简单排列图：单个个案的值】对话框，见图 19-59。例题数据文件为 data19-17，图为汽车空调蒸发器故障数据，见图 19-60。将次数变量送入【值】框中，在【类别标签】栏选择【变量】，将问题分类变量送入【类别标签】栏的矩形框中，单击【确定】按钮，得到图 19-60。从图中可以了解到丢失螺钉是造成故障的最主要原因。

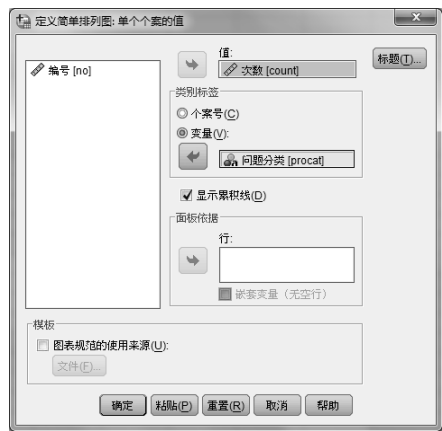


图 19-59 【定义简单排列图：单个个案的值】对话框

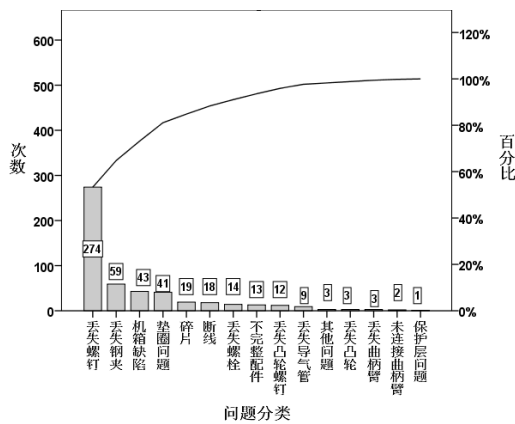


图 19-60 例图

19.7.3 堆栈帕累托图

1. 以个案组计数统计模式生成的堆栈帕累托图

在【帕累托图】对话框中选择【堆积面积图】和【个案组的计数或总和】项，单击【定义】按钮，展开【定义堆积排列图：个案组的计数或和】对话框，见图 19-61。例图为各洲具有加工制造业工厂数量帕累托图，见图 19-62。数据文件为 data19-18。主要操作步骤是：在【条的表征】框中选择【变量和】项；将 count 变量选入此框；选择洲 cont 变量进入【类别轴】框；

选择 cat 分类变量进入【定义堆栈】框。单击【确定】按钮生成图 19-62（堆栈条形中的数字已经删去）。



图 19-61 【定义堆积排列图：个案组的计数或和】对话框

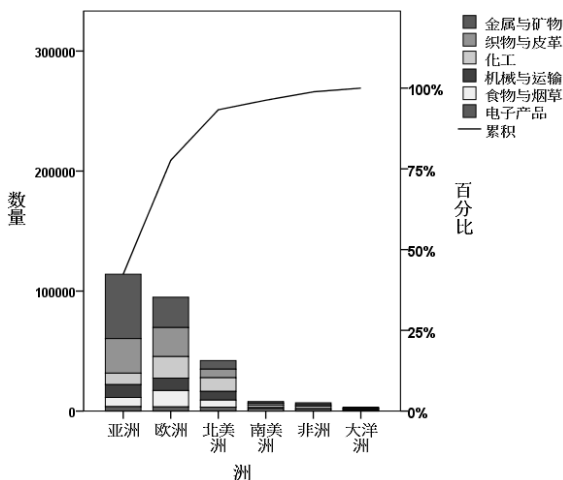


图 19-62 例图

从图 19-62 可以看出几乎 80% 的 100 人以上的工厂集中在亚洲和欧洲，以金属与矿物、织物与皮革为最多。

## 2. 以个案值统计模式生成的堆栈帕累托图

data19-19 是不同性别各年龄段司机百万公里伤亡及非伤亡事故的统计数据。在【帕累托图】对话框中选择【堆积面积图】和【个案值】选项，单击【定义】按钮，展开【定义堆积排列图：单个个案的值】对话框，见图 19-63，将男女司机伤亡事故数变量、非伤亡事故数变量送到【值】框中；类别标签选择变量选项，将年龄段变量送入变量下的框中。单击【确定】按钮，生成各年龄段司机交通事故例数帕累托图，见图 19-64。

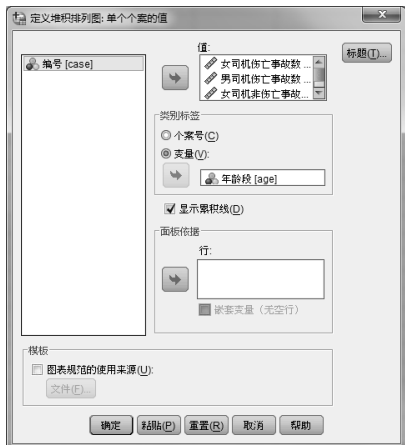


图 19-63 【定义堆积排列图：单个个案的值】对话框

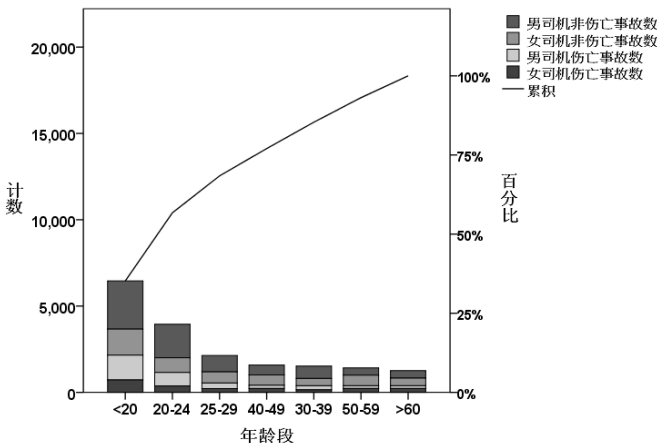


图 19-64 例图

从图中可以看到，非伤亡交通事故与伤亡交通事故总和在 30 岁之前占了近 80%；交通事故随年龄增长逐渐减少；非伤亡事故占大多数。

# 19.8 控 制 图

控制图（Control Charts）又称管理图，它是用于分析和判断生产工序是否处于稳定状态所使用的一种带有控制界限的统计图。虽然它始于产品质量的控制，但以后推广到生产领域以外的许多方面，如医学、金融等领域。控制图大致分为两类：一类是计量值控制图，另一类是计数值控制图。在实际应用中，这两类控制图常常是组合使用的。

## 19.8.1 选择控制图类型

按【分析→质量控制→控制图】顺序单击菜单项，打开【控制图】对话框，见图 19-65。

### 1. 控制图图式

- (1)【X 条形图、R 图和 s 图】。包括两种组合控制图，X-Bar、R 平均值-极差组合控制图和 X-Bar、s 平均值-标准差组合控制图。
- (2)【个体，移动全距】。单值-移动极差组合控制图。
- (3)【p, np】。包括 p 不合格品率和 np 不合格品数两种控制图。
- (4)【c, u】。包括 c 缺陷数控制图和 u 单位缺陷数控制图。

### 2. 数据组织选择

- (1)【个案为单元】。观测量组结构数据选择此项。如数据文件 data19-19 的数据结构。
- (2)【个案为子组】。变量组结构数据选择此项。如数据文件 data19-21 的数据结构。



图 19-65 【控制图】对话框

## 19.8.2 平均值、极差、标准差控制图

### 1. 个案为单元的平均值、极差、标准差控制图

在【控制图】对话框中选择【X 条形图、R 图和 s 图】和【个案为单元】项，单击【定义】按钮，打开相应对话框，见图 19-66。根据数据文件 data19-20 中的数据作图，该数据是某厂 1988 年 6 月电解工序三班的电解效率数据。



图 19-66 【X 条形图、R 图、s 图：个案为单元】对话框

- (1)【过程度量】。指定控制图主要描述的变量。本例中指定“电解效率【eec】”变量送入该框。
- (2)【定义子组】框指定一个分类变量，在控制图中作为横轴变量。本例选定“日期【date】”变量送入该框。
- (3)【图表】栏有两种组合：【X 条形图使用范围】和【X 条形图使用标准差】。这两个组合控制图的使用区别在于，前者用于细分组中样本数量较小的资料，后者用于定义子组变量的各分类中样本数量较大（大于 10）的资料。本例选定【X 条形图使用范围】项。

(4)单击【选项】按钮,打开【X 条形图、R、s: 选项】对话框,见图 19-67。

①【Sigma 的数目】。选择中心线上、下的标准差数值,默认值为“3”。

②【最小子组大小】。指定分类变量各类中最小样本数,默认值为“2”。

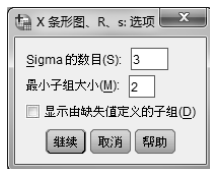


图 19-67 【X 条形图、R、s: 选项】对话框

③【显示由缺失值定义的子组】。在图中缺失值作为一类显示。

单击【确定】按钮,作图输出,见图 19-68。

图 19-68(a)所示为每日三班电解工序的电解效率的平均值-极差控制图。中间横线位置是均值 94.739; UCL 是控制管理的上限 100.043; 控制管理下限 LCL 为 89.435。

图 19-68(b)所示是电解效率范围-极差控制图。这里的平均值是每天电解效率的范围的平均值。范围即每日三班中电解效率的最大值减最小值。范围的平均值是 5.183, 管理控制的上限为 13.345, 三班的效率最大差不能超过此值, 最小为 0。

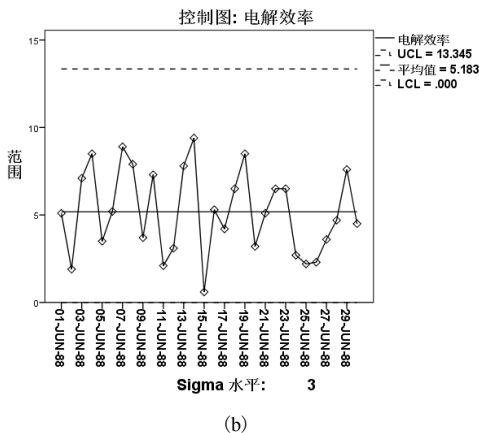
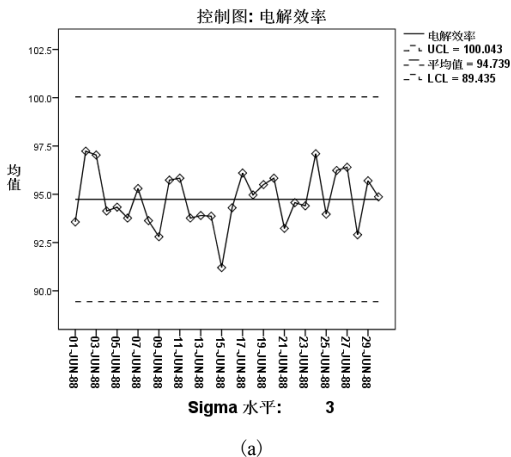


图 19-68 例图

## 2. 个案为子组的平均值、极差、标准差控制图

数据文件 data19-25 是某轧钢厂对 6mm 钢板的测试记录数据。变量 case 是样品编号, t1~t5 是对每个样品进行 5 次测试所得的结果数据。



图 19-69 【X 条形图、R 图、s 图: 个案为子组】对话框

在【控制图】对话框中选择【X 条形图、R 图和 s 图】和【个案为子组】项, 单击【定义】按钮, 打开【X 条形图、R 图、s 图: 个案为子组】对话框, 见图 19-69。

(1)【样本】列表。样品测定, 至少选定 2 个或 2 个以上的数值型变量, 本例选择了 t1~t5 共 5 个变量 (对每个样品的 5 次测量值变量)。

(2)【标注子组】框。细分组标识。本例选用 case 变量作为细分组标识变量, 在横轴上显示 19 个样品的均值或 5 次测量的范围 (最大值减最小值)。

图 19-70(a)所示为 6mm±0.4mm 厚度钢板平均

值控制图。测量的平均值是 5.9906，UCL 控制管理的上限值为 6.2836，下限为 5.6976。图中每个点的纵坐标为每个样品 5 次测量的平均值。

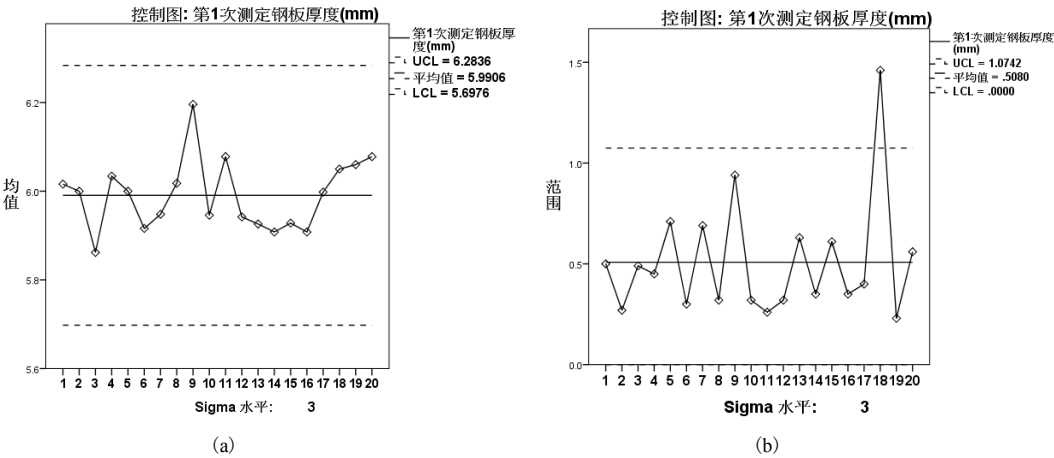


图 19-70 例图

图 19-70(b)所示为极差控制图。图中每个点纵坐标为每个样品 5 次测量值中的最大值最小值的范围值，也称极差值。这些范围的平均值是中间的横线纵坐标，为 0.5080，测量范围的控制管理上限为 1.0742，下限为 0。

这两个图的题目有误，不是“第 1 次……”，可以在编辑图形时改变。这里原样给出，为了方便读者对照。

19.8.3 单值-移动极差控制图

数据文件 data19-22 为混凝土塌落度数据。

在【控制图】对话框中选择【个体，移动全距】和【个案为单元】项，单击【定义】按钮，打开【个体和移动全距】对话框，见图 19-71。



图 19-71 【个体和移动全距】对话框

- (1) 【过程度量】框。选择 value 变量作为作图变量。
- (2) 【标注子组】框。选定 no 变量为细分组的标识变量，即横轴变量。

- (3) 【图表】栏。
  - ① 【个体和移动全距】。绘制单值-移动极差控制图，本例选定该选项。
  - ② 【个体】。绘制单值控制图。
  - ③ 【跨度】。指定计算控制极限时所使用的个案数量，以及用于计算移动范围的时间单位跨度。默认值为“2”。可以指定 2~100 之间的整数。

输出例图见图 19-72。图 19-72(a) 为个体控制图。每个点是一个观测值。中间直线为这些观测值的平均值 7.104；控制管理上限为 9.019，下限为 5.188，也就是混凝土塌落度应该控制在这个区间。

图 19-72(b) 为移动全距图。愿望跨度是 2，第一个点在横坐标 2 处，其值为第二个点值减第一点值之差，第二个点在横坐标为 3 处，其值为第三个点值减第二个点值之差，依次类推。



移动差值的平均值为 0.721，控制管理上限为 2.355，下限为 0。两次处理之间的差值应该控制在这个范围内。

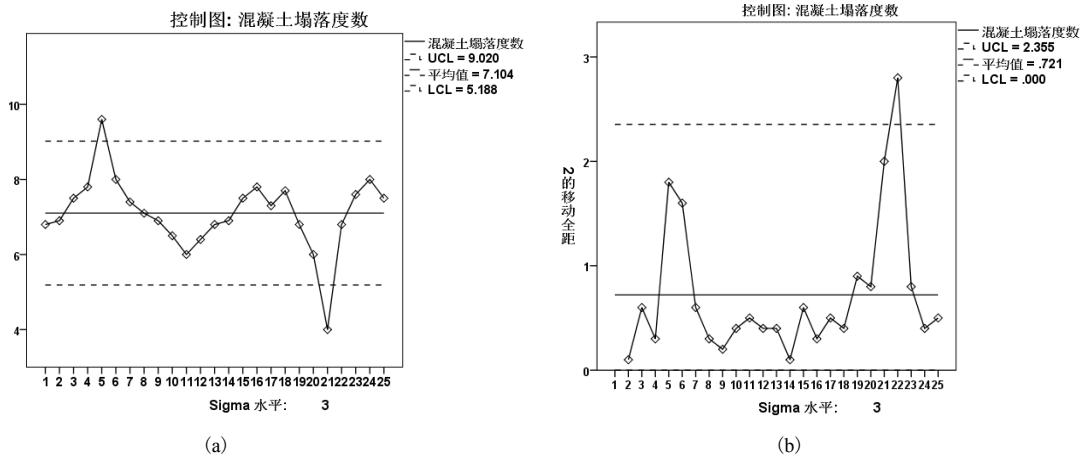


图 19-72 例图

## 19.8.4 不合格品率、不合格品数控制图

### 1. 个案为单元的不合格品率、不合格品数控制图

在【控制图】对话框中选择【p, np】和【个案为单元】项，单击【定义】按钮，打开【p、np：个案为单元】对话框，见图 19-73。

例题数据文件为 data19-23 小螺钉检测数据-1.sav。case 为样品分组变量；products 为样品是否合格的数值变量，值为 0 为不合格，值为 1 表示合格。

(1) 【特征】框。指定作图变量，必须是数值型。本例选择 products 变量作为作图变量。

(2) 【计数值】栏。变量值计数方式。

① 【不符合（应为不合格）】。指定计算不合格产品。

② 【符合（应为合格）】。指定计算合格产品。

③ 【值】框。变量值属性，在数据文件 data19-23 中，如果选择计算不合格产品数量，选择【不合格】项，并在【值】框中输入“0”。这里输入的变量值应与作图变量中的变量值类型相同。

(3) 【定义子组】。细分组标识变量。本例选用 case 变量。

(4) 在【图表】栏选择图形描述模式。

① 【p（比例不符合）】。作不合格品率控制图。

② 【np（数目不符合）】。作不合格品数控制图。本例选择此项。

不论在【计数值】框选择了计算【不符合（应为不合格）】还是【符合（应为合格）】选项，最后生成的图形都为不合格品数或不合格品率的控制图。

图 19-74 所示为不合格品数控制图，中间的横线是中心线，各点代表横轴坐标表示的组内不合格品数量。

### 2. 个案为子组的不合格品率、不合格品数控制图

在【控制图】对话框中选择【p, np】和【个案为子组】项，单击【定义】按钮，打开【p、np：个案为子组】对话框，见图 19-75。本例使用数据文件 data19-26 和 data19-27，生成的例图见图 19-76。



图 19-73 【p、np：个案为单元】对话框

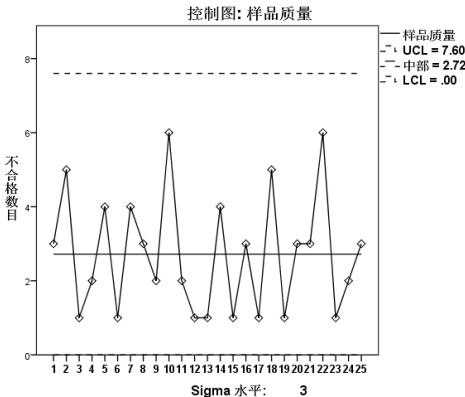


图 19-74 例图



图 19-75 【p、np：个案为子组】对话框

数据文件 data19-26 为某构件厂产品质量数据。变量 no 是样品分组编号，变量 sam 是样品数，变量 unq 是不合格数。每个分组样品数均为 500。

数据文件 data19-27 为抽样数不等的小螺丝检测数据。变量名及含义与 data19-26 相同，只是每个分组样品数不同。

(1) 【数目不符合】框。样本不合格的数量。均将变量 unq 移入该框。

(2) 【标注子数】框。即细分组标识。均将变量 no 样品分组移入该框。

(3) 【样本尺寸】栏。

① 【常量】。细分组样本数量恒定，每个细分组样本量相同，选择此项并在框中输入样本数。在数据文件 data19-26 中，由于每个细分组的样本数都是 500，故在此框内输入“500”。

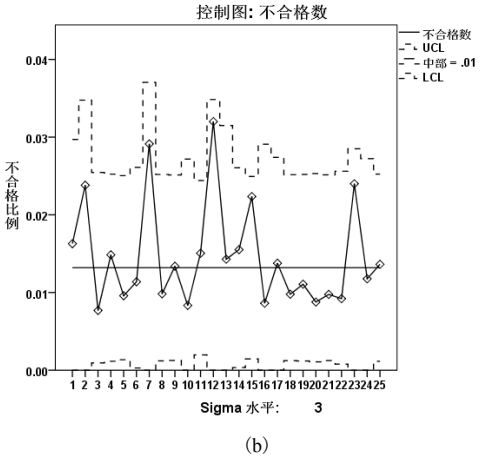
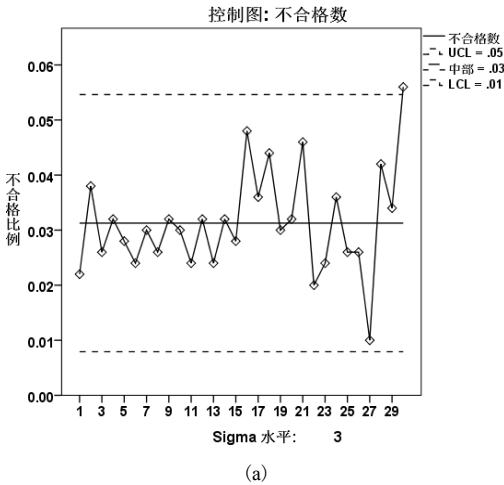


图 19-76 例图

② 【变量】。确定样本数量，无论细分组样本数目是否相同，都可以通过表示每个细分组

样本数量的变量来说明细分组样本数目。数据文件 data19-27 中的 sam 变量就是这样一个变量，所以用 sam 变量确定细分组样本数量。样本尺寸选择此项。

使用数据文件 data19-26，在【数目不符合】框中选中 unq 变量；选用 no 变量作为细分组标识的变量；在【样本尺寸】栏中选择【常量】项，并输入 500；最后在【图表】栏中选择【p(比例不符合)】项，生成某构件厂产品不合格品率控制图，见图 19-76(a)。

使用数据文件 data19-27，在【数目不符合】框中选中 unq 变量；选用 no 变量作为细分组标识变量；在【样本尺寸】栏中选择【变量】项，在该框内输入 sam 变量；最后在【图表】栏中选择【p(比例不符合)】项，生某种小螺钉不合格品率控制图，见图 19-76(b)。

由于数据文件 data19-27 中各分组中的样品数不同，所以不合格比例控制的上、下限不在一条水平线上。

### 19.8.5 变量缺陷数、单位缺陷数控制图

在【控制图】对话框中选择【c, u】和【个案为单元】项，单击【定义】按钮，打开相应对话框，见图 19-77。本小节使用数据文件 data19-24，为某医院每周危急手术例数。例图 19-78 为某医院每月出现危急外科手术的缺陷数控制图。



图 19-77 【c, u: 个案为单元】对话框

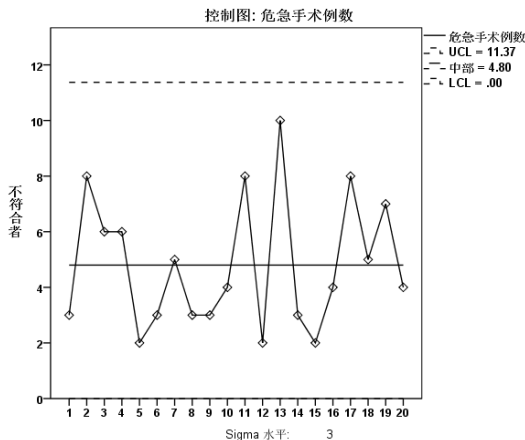


图 19-78 例图

主要操作步骤：选择 aes 变量作为被测对象送入【特征】框；选用 week 为细分组标识的变量送入【定义子组】框。【图表】栏有两个选项：

- ① 【u (每个单元的不符合数)】。生成单位缺陷数控制图。
- ② 【c (不符合的数目)】。生成缺陷数控制图。本例选择此项。

## 习 题 19

1. 绘制统计图形有哪些基本要求？
2. 使用数据文件 data19-28，绘制以下图形：
  - (1) 男性和女性期望寿命对比条图；
  - (2) 不同气候地区的国家数量图；
  - (3) 不同地区平均国民生产总值交互图；
  - (4) 世界上不同宗教所占百分比饼图。

# 第 20 章 编辑统计图形

## 20.1 认识图形组成

系统生成“普通”统计图和交互统计图，它们各部分的名称基本相同。这里介绍一些主要的名称。

图形边框划定了图形区域，见图 20-1。图形外框包括整个图形。数据边框包括坐标轴标题、坐标轴上的数值标注以及图形。图形内框仅包括图形。图形外框和数据边框可以拖曳移动点，改变图形相应部分的大小。文本可以不同形式出现在图形中，可以是图形标题、脚注、轴标题、数值标签、注释等。

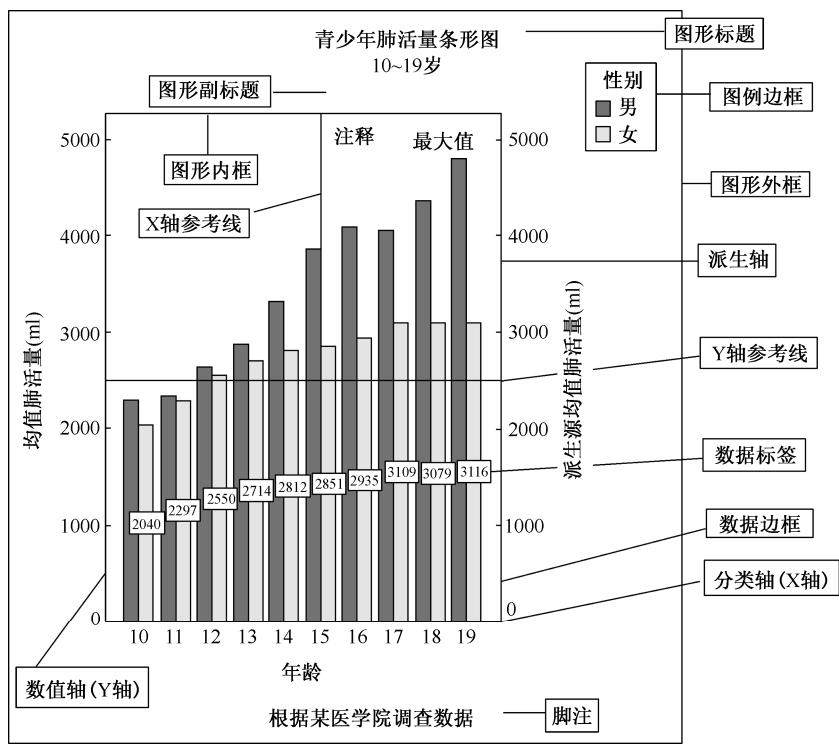


图 20-1 图形元素

其他图形组成见图 20-2。

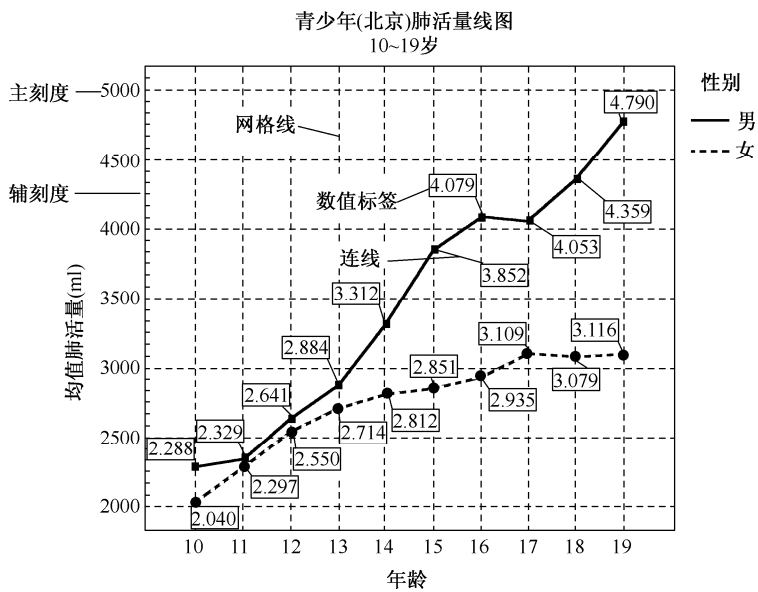


图 20-2 图形组成说明

## 20.2 编辑平面统计图

### 20.2.1 图形编辑途径

在输出观察窗中产生图形后,为了进一步探查数据或增强视觉效果,需要在【图表编辑器】窗口编辑所生成的图形。

#### 1. 编辑图形的三种途径

要编辑生成的图形,基本操作是双击它,进入图形编辑状态,即在【图表编辑器】窗口显示待编辑的图形,如图 20-3 所示。同时打开【属性】窗口,见图 20-4。

图 20-3 所示的图形编辑窗标题栏下面,自上至下分别为功能菜单、编辑工具、选择工具、元素工具和格式工具。窗口右下角还有状态栏。是否在窗口中显示这些工具,可以在【查看】菜单项中选择。

(1) 在【图表编辑器】窗口使用菜单项和工具栏中的工具对图形进行编辑是对图形编辑的第一种途径。这里许多菜单中的命令已经在前面的章节中介绍过。

(2) 在打开编辑器的同时,如果没有打开【属性】窗口,可以按【编辑→属性】顺序单击菜单项,打开【属性】窗口;也可以在外框内空白处右击,选择右键菜单的第一项【属性窗口】,打开【属性】窗口。在【图表编辑器】窗口的图形上选择了一个要修改的图形元素后,【属性】窗口的内容发生相应的改变。各选项卡中的编辑方法与选中的图形元素对应。使用【属性】窗口中的各种选项卡中的功能是编辑图形的第二种途径。

(3) 右击待编辑的图形元素时展开右键菜单,其中包括各种编辑该图形元素和有关的元素组的功能。图形不同、选择的图形元素不同,右键菜单的内容也各不相同,选择其中的功能项,也可以对图形元素进行具体的编辑。这是编辑图形的第三种途径。

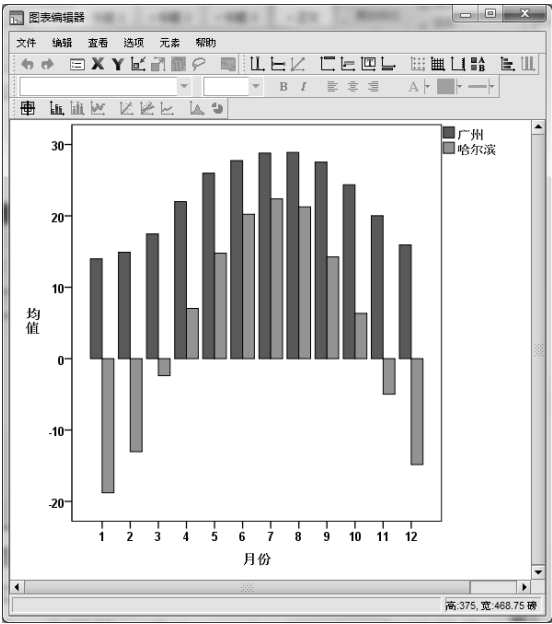


图 20-3 【图表编辑器】窗口

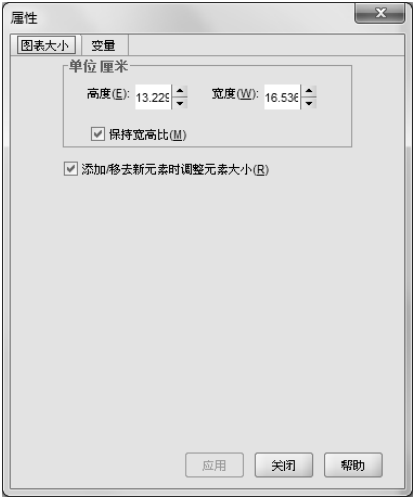


图 20-4 【属性】窗口

这三种途径是相通的，但都必须在【图表编辑器】窗口中才能实现编辑功能。本节主要介绍与图形编辑有关的命令、工具、属性窗和右键菜单的操作。

2. 选择编辑对象

(1) 图形元素的选择。要对图形元素进行编辑，必须首先选择它。要编辑的图形元素有时是单个元素，如选择坐标轴；有时是一组，如轴上刻度的标签。被选择的图形元素被彩色框框住。选择方法很多，下面介绍几种常用方法：

- ① 单击选择。例如，选择坐标轴，选择饼图的所有扇面，选择图形的数据区。
- ② 双击选择。往往用于选择并列元素组中的一组元素，如选择双变量条形图中的一个变量的一组条。
- ③ 右键菜单选择。用于指定元素组中的一个成员。例如，要选择条形图中的一个条，可右击其中一条，在右键菜单中选择【Select→此条】。

④ 用套索选择几个图形元素。例如，在散点图中选择离群点，可以单击套索工具，用套索光标对要选择的对象画封闭曲线。

(2) 文字的选择与移动、放缩。选择文字的方法与图形元素的选择方法相同。选择后，文字四周出现带有 8 个方块的框，见图 20-5。将鼠标指针置于方块上，按住左键，鼠标光标变成双向箭头，拖拽这个光标，可以放缩套住的文字；当鼠标光标置于框的边缘，鼠标光标变成 4 个箭头，可以按住左键，移动图形元素到新位置。



在如图 20-4 所示窗口的【图表大小】选项卡中可以精确地放缩图形大小，单击【高度】、【宽度】旁的上下箭头，图 20-5 选择与移动、缩放对象 改变图形外框的高度和宽度。要想保持高宽比不变，按比例放缩，选择【保持宽高比】。缩放外框以内的各图形元素时文字部分，如主副标题、轴标题、轴刻度标签、图例不变化。还可以选择【添加/移去新元素时调整元素大小】。

## 20.2.2 改变图形构成

为了说明编辑方法,先作条形图。数据文件 data20-01 中是 12 个城市 1985—1994 年各月的气温数据。单击【图形→旧对话框→条形图】菜单项,在【条形图】窗口选择【复式条形图】,单击【定义】按钮,打开【定义复式条形图:各个变量的摘要】对话框。在源变量框中选择月份作为 X 轴变量送入【类别轴】框;把广州和哈尔滨的气温广州、哈尔滨两个变量送入【条的表征】框。

单击【标题】按钮,输入图形标题,第一行:“月均气温比较”;第二行:“1985—1994”。单击【继续】按钮,在主对话框中单击【确定】按钮,生成图形如图 20-6(a)所示。

### 1. 图形转换

图形转换必须要有充足的数据。系统自动识别可以转换的图形,在【属性】窗口【变量】选项卡的【元素类型】下拉列表上加黑的图形就是当前图形可以转换成的目标图形。将标记状态的条形图转换为内插线图和散点图操作方法是:双击该条形图,在【属性】窗口【变量】选项卡的【元素类型】下拉列表中选择【内插线图】,单击【应用】按钮,图形变换成图 20-6(b)所示的线图;在下拉列表中选择【标记】,单击【应用】按钮,则图形转换成图 20-6(c)所示的形状。每选择一种,单击一次【应用】按钮。

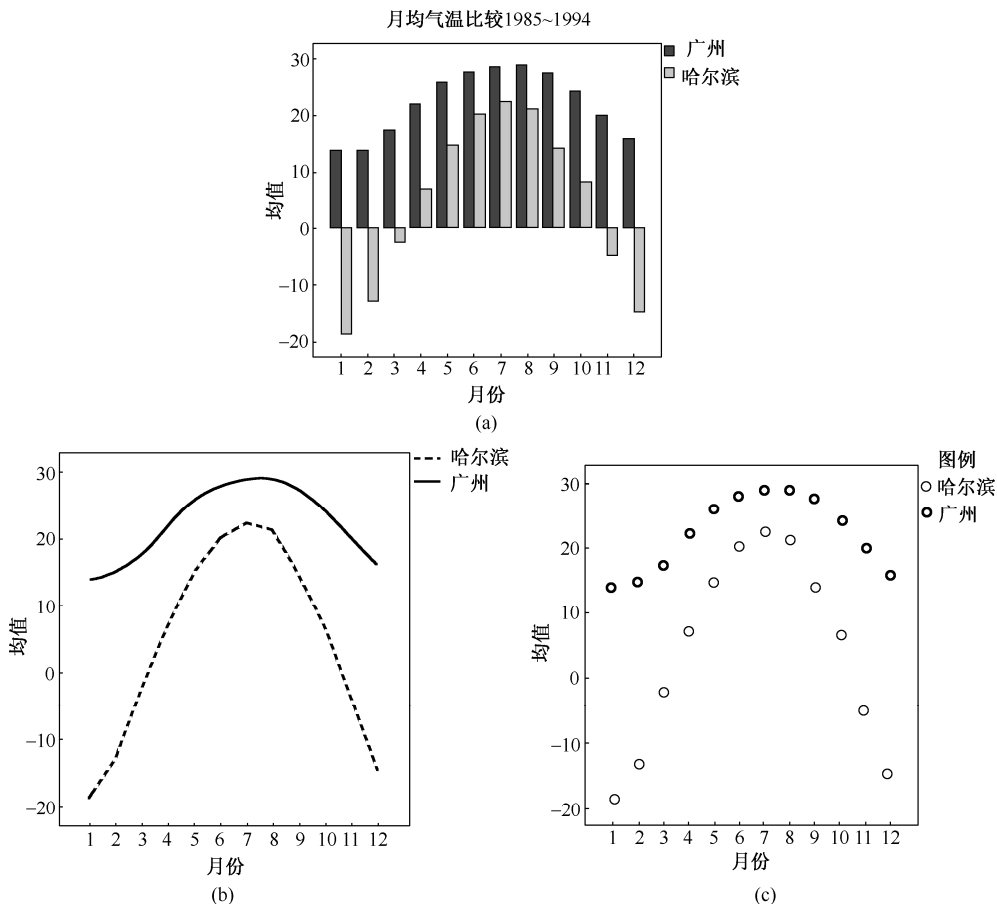


图 20-6 图形转换

注意：变化后的图形类型不一定能很好地表达数据，不一定能方便观察。例如，本例若要转换成饼图，就不易直接观察。所以要注意选择最后的转换结果。

2. 图形转置

对有 X、Y 轴的平面图形，可以进行转置，即把直角坐标系旋转 90°。这与改变源变量与自变量的角色还是有区别的。例如，对曲线图形来说，后者需要重新进行拟合。

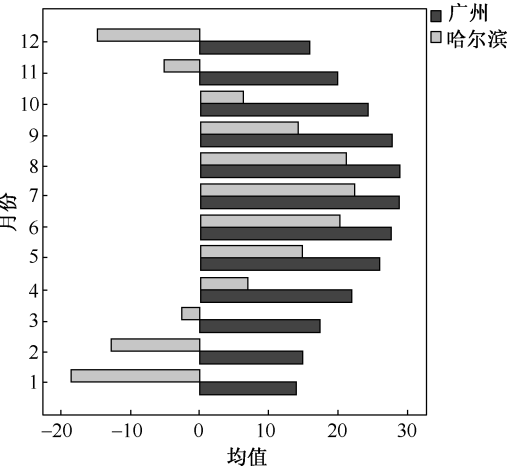


图 20-7 转置结果

选择右键菜单中的【隐藏数值标签】，值标签消失。

选择值标签，【属性】窗口会增加【数据值标签】选项卡，其中【显示】框中是已经显示的标签，【不显示】框中是还没有显示但可以加上的标签。例如，选择“月份”送入【显示】框，单击【应用】按钮，哈尔滨的气温条上增加了数据值标签和月份值，见图 20-8(b)。增加了【数据值标签】选项卡的【属性】窗口，如图 20-9 所示。

图形转置的方法很简单，在【图表编辑器】窗口右击，在右键菜单中选择【变换图表】，图形顺时针旋转，得到转置后的图形，见图 20-7。

3. 在图形中增加值值标签

用户可以显示条形图中的条、饼图中的扇、线图上的点、箱线图的中线所代表的数值、百分比，或散点图和箱线图各个观测量的数值。

首先选择要显示的数值的图例，在图 20-8(a)所示图例中选择“哈尔滨”，在右键菜单中选择【显示数值标签】，所有的哈尔滨平均气温数值标出；

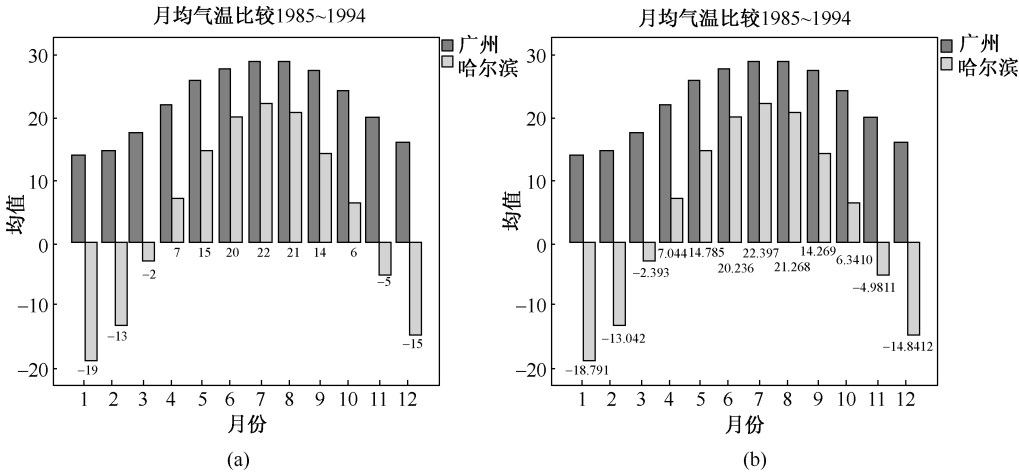


图 20-8 数值标签

单击工具栏中的“”工具，光标变成“”形状时，表示激活数据识别模式，单击某个图形中的条、点、线等即可显示/隐藏数值标签。这个工具的方便之处在于可以逐个增加标签，也可以逐个隐藏已经加上的标签。

在【标签位置】栏可以选择【自动】、【手动】、【自定义】之一以调整标签的位置。选择【手



动】调整标签位置，可以用鼠标拖拽已经加在图中的标签到任何位置；选择【自定义】时可以在下面的矩阵框中选择标签相当于条的位置。

4. 增加其他图形组成

图形生成后可以对图形继续修饰。例如，增加对图形的解释，对一些变量或数值注释，画出参考线、拟合线等。图 20-10 所示是右键菜单，增加图形元素的选择项有【添加 X 轴参考线】、【添加 Y 轴参考线】、【添加标题】、【添加注释】(内框内、数据区中)、【添加文本框】(外框内、内框外)、【添加注脚】。

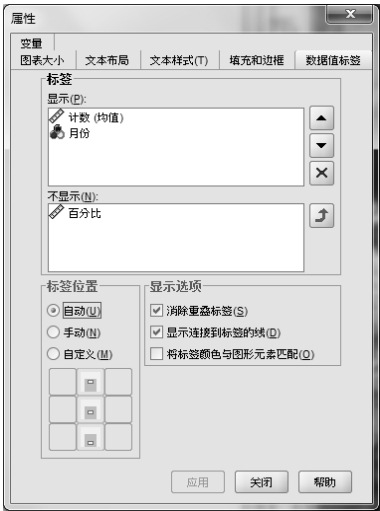


图 20-9 【数据值标签】选项卡



图 20-10 右键菜单

图 20-11 所示是添加图形元素 Y 轴参考线(Y 轴 10° 参考线)、注释(最高平均气温、最低平均气温)、注脚(2014 年 5 月制图)的结果。

这些新加入的图形元素与生成的图形元素一样，单击、双击可以选择这些元素进行编辑，可以移动位置，还可以对文字元素改变字体、字号等。

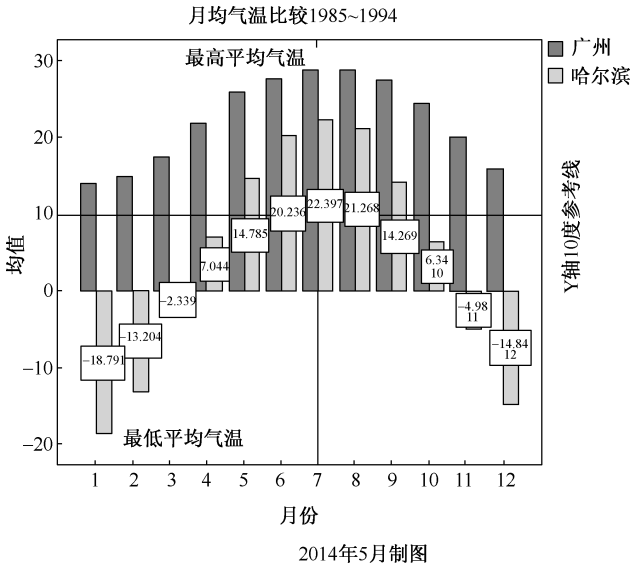


图 20-11 添加新图形元素的条形图

5. 显示派生轴、图例、线图标记点

派生轴是为了看图方便在原始轴对面产生的刻度、标注与原轴不完全一样的轴线。在右键菜单中选择【显示派生轴】、【隐藏派生轴】可对派生轴显示与隐藏进行切换。如图 20-12 中右侧有刻度的竖线就是派生的 Y 轴。派生轴上的刻度是可以通过设置改变的,不一定与原轴相同。

在原图中有图例。在右键菜单中选择【隐藏图注】、【显示图注】可以进行显示与隐藏图例的切换。

线标记是在线图上对应横轴各刻度的点的标记。如图 20-12 中曲线上的各点标记。编辑线图时,右键菜单中可选择【显示线标记】、【隐藏线图标记】对线上标记进行显示或隐藏的切换。

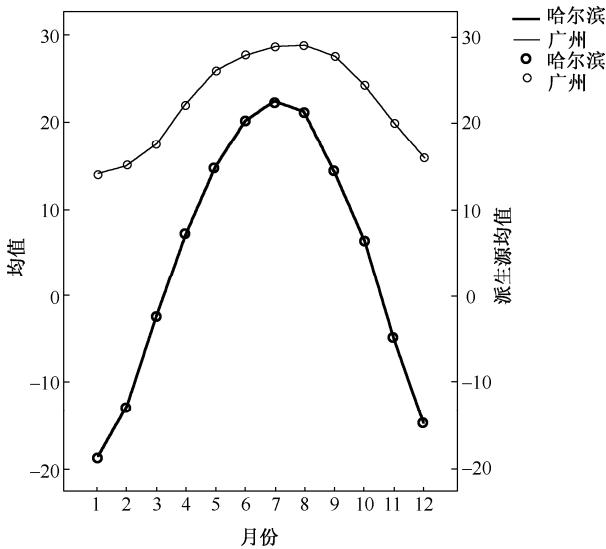


图 20-12 派生轴、图注和标记点

6. 改变分类在属性窗【分类】选项卡中进行

例如,对条形图、散点图等改变图中分类变量各类的顺序、增加或减少某一类的条或点,对饼图减少增加某一类的扇面等。

【例 1】以数据文件 data20-01 为例,建立编辑图形的概念,认识常用编辑功能的用法及其效果。

1. 建立条形图

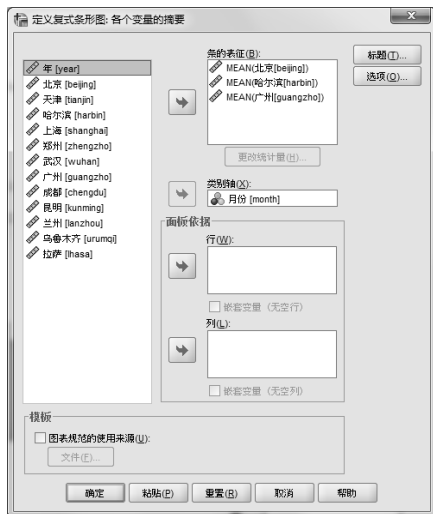
从图形菜单中单击【图形→旧对话框→条形图】打开如图 20-13(a)所示【条形图】对话框。在对话框中选择【复式条形图】,在【图表中数据为】栏选择各个变量的摘要。单击【定义】按钮。

在如图 20-13(b)所示的【定义复式条形图:各个变量的摘要】对话框中,选择“北京”、“哈尔滨”、“广州”变量送入【条的表征】框,“月份”变量送入类别轴,单击【确定】按钮,生成条形图。双击该图进入编辑状态,见图 20-13(c),图 20-13(d)所示是对应的图形【属性】窗口。

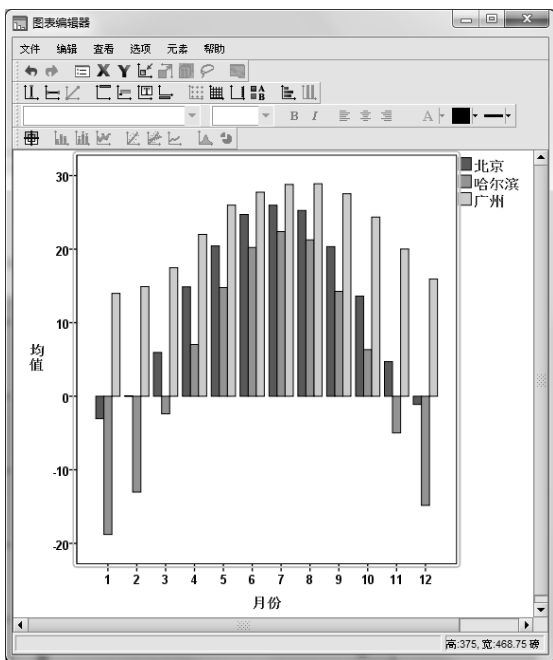
在【属性】窗口单击【图表大小】选项卡,见图 20-13(d),可以精细地调整图形外框的尺寸。选择【保持宽高比】选择项,调整长、宽中的一个尺寸,另一个尺寸会自动根据比例变化,不用再手动调整;选择【添加/移去新元素时调整元素大小】是自动完成的。如果不满意还可以手动调整。



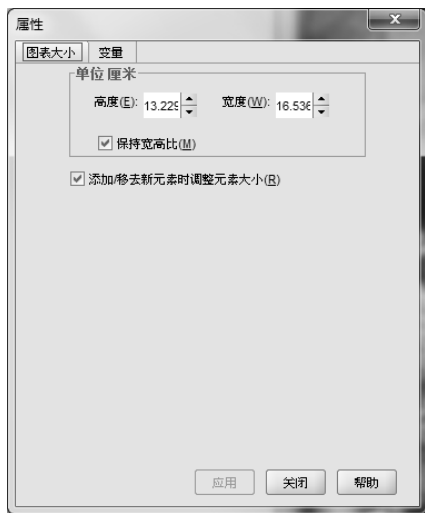
(a)



(b)



(c)



(d)

图 20-13 生成条形图与进入编辑器的操作示意图

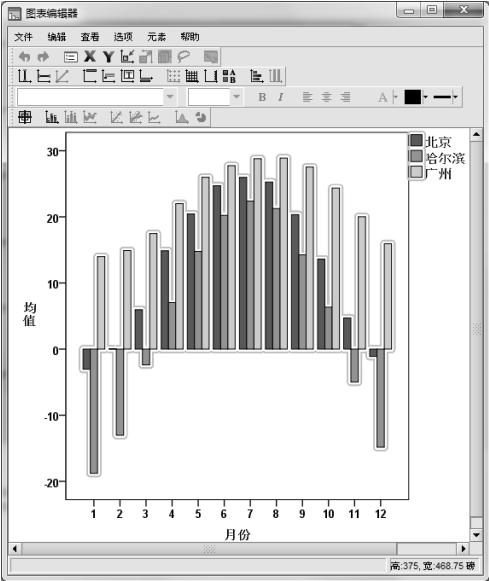
## 2. 编辑条形图

当单击一个条时，选择所有的条，如图 20-14(a)所示。【属性】窗口同时发生变化，如图 20-14(b)所示，这里显示【类别】选项卡。

(1) 在【变量】下拉列表中可以指定变量，也可以指定【图例】，指定一个设置一个，单击【应用】按钮实现一个。图 20-14(a)所示为对图例所列城市分类的操作。图 20-14(b)所示为对变量北京的操作，去掉“北京”的图条：在【顺序】框中选择“北京”，单击栏右侧的叉子图标，“北京”被移到【已删除】框内，见图 20-14(c)，输出条形图中就没有表示北京变量的 12 个月平均气温的条了，见图 20-14(c)所示。

(2) 在【折叠(汇总)小于以下值的类别】框中输入百分比值，将小于设置数值的元素合并为一类。例如输入“10”，凡是图中数值总和小于 10%的分类，合并为默认名为“其他”的新分类显示在【顺序】框中。这个功能用于在条形图重点不突出时合并一些类。

(3) 【类别】栏显示分类变量的各类排序的方法和方向。



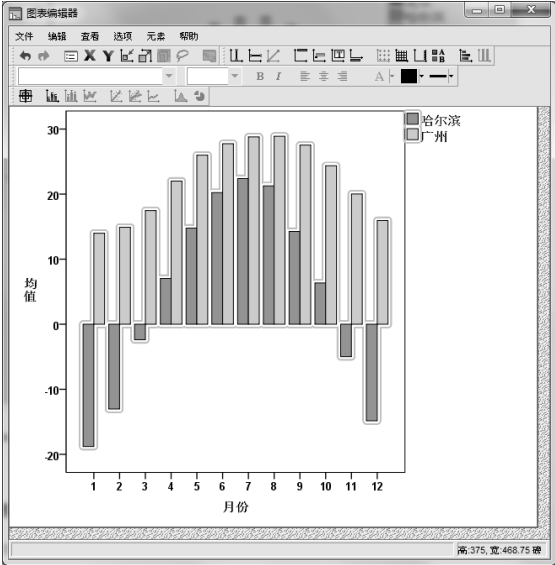
(a)



(b)



(c)



(d)

图 20-14 【类别】选项卡对分类变量的设置

- ① 【排序依据】下拉列表中选择分类按【值】、【标签】、【统计】、【自定义】的顺序排序，排序结果显示在【顺序】框中。
- ② 选择前 3 种排序分类值的方法，需要在【方向】下拉列表中选择【升序】或【降序】排序。
- ③ 【顺序】框中显示分类轴上显示的分类值。如果选择了【自定义】，可通过上、下箭头

按钮调整分类值的顺序。选中某分类值，单击叉子按钮将其移到【已删除】框中，分类轴上不再显示该分类值的图形。

④ 【已删除】框中是被剔除的分类变量。选中一个，单击向上按钮，该变量送回【顺序】框。

(4) 分类轴两侧留白选择项【上边距(%)】、【下边距(%)】分别表示在分类轴上面、下面留出的空间占整个分类轴的百分比。

图 20-15(a)所示是独联体中 4 个国家失业月数据条形图，上图为原图，下图为去掉 1 个国家数据条的结果。

【类别】选项卡的功能对饼图(也称圆图)的作用更明显。图 20-15(b)中的上图是原图，下图是将小于 8% 的扇面合并，按统计量排序 1 月、2 月的扇面合并成其他类，数值是 2 个月的数值相加，为 15.11%。

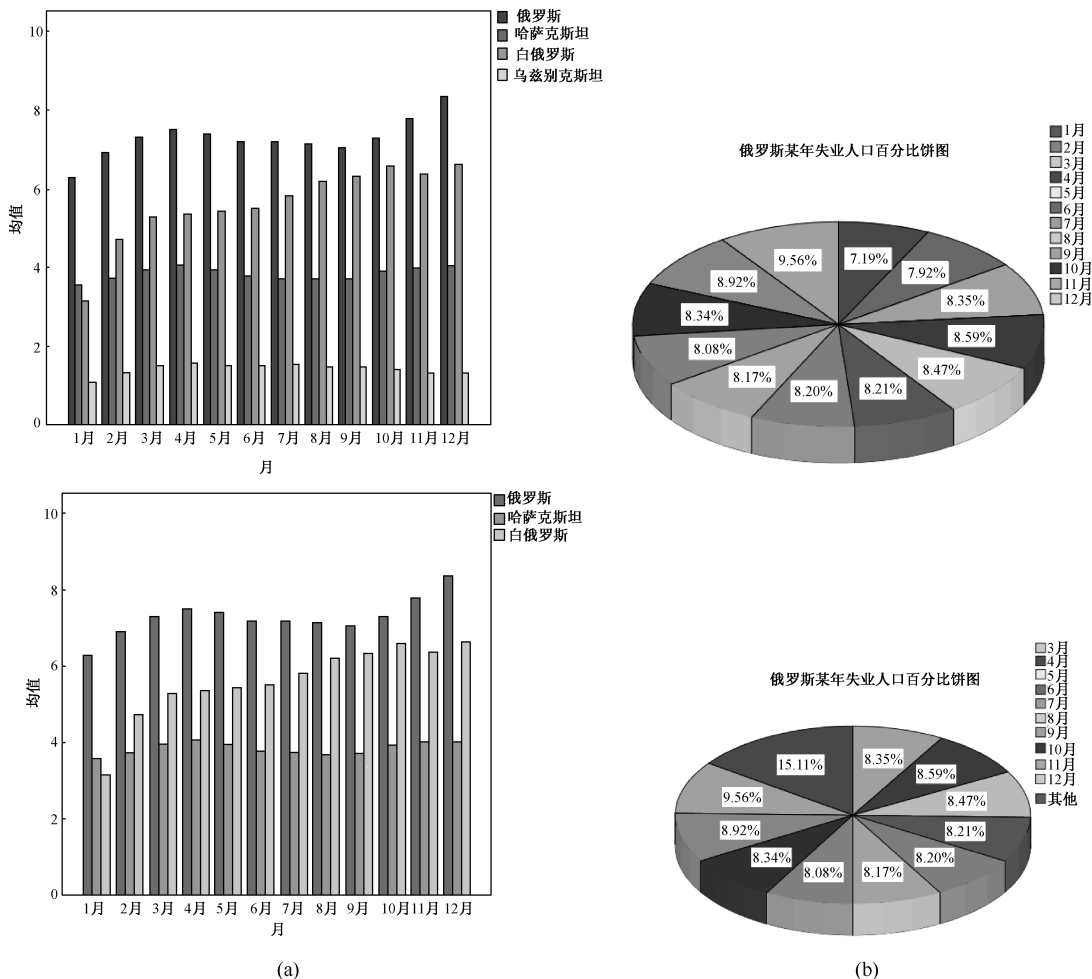


图 20-15 原图与编辑后的效果图比较

### 20.2.3 图形与文字修饰

对图形的修饰在【属性】窗口中的【填充和边框】选项卡中进行，当选择了条形图的条、饼图的扇形等时，【属性】窗口自动变为图 20-16(a)所示窗口；当选择了文字元素，如数值标

签、标题、注脚等时，【属性】窗口自动变为图 20-16(b)所示带有【文本样式】选项卡的窗口。在右键菜单中选择【显示数据标签】后，编辑数据标签的【文本布局】选项卡出现在【属性】窗口中，见图 20-16(c)。

如果选择了要修饰的图形元素或文字，没有出现带有相应功能的【属性】窗口，按【编辑→属性】顺序单击菜单项，打开【属性】窗口。

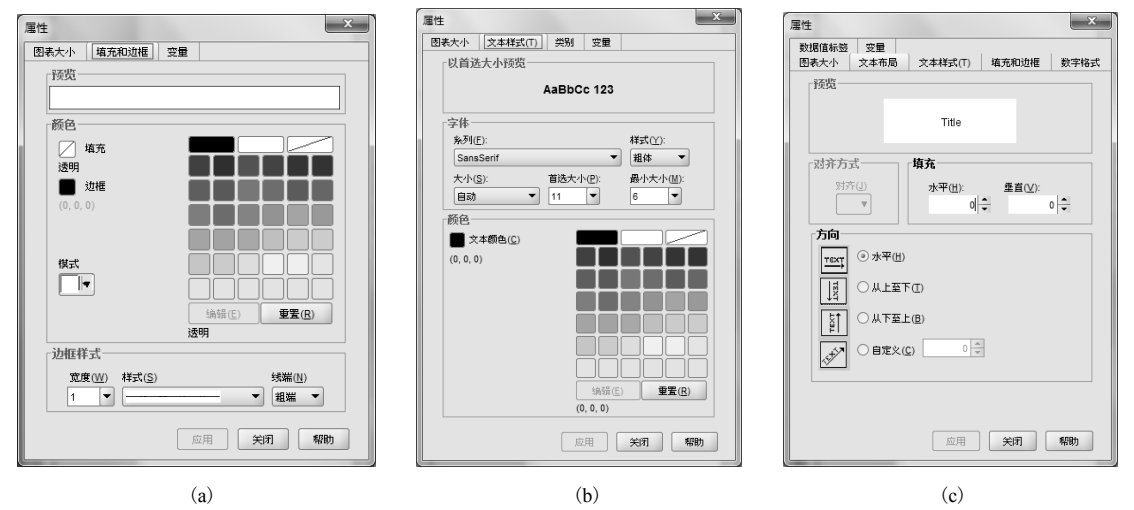


图 20-16 修饰图形元素、文字的【属性】对话框

1. 填充和边框

填充功能是对图形中的整体或选中的区域进行填色或增加底纹。边框是对选定的区域增加线框，改变边框的线型、粗细、颜色。被选定区域可以包括全部图形、图形内框区、图例框、文本框、注释框等，还包括条图、面积图、极差图、饼图、箱线图、误差条图、直方图等。

- (1) 【预览】框显示单击【应用】按钮后实现的填充颜色、底纹、框线的效果。
- (2) 【颜色】栏。【填充】框中显示填充的颜色，【边框】框中显示边框颜色，【模式】下拉菜单中选择要填充的底纹。在调色板中，选择[填充黑色]，选择[填充白色]，选择[填充透明色]。内框内背景色在创建图形后显示为灰色，选择[透明的填充]效果较好。
- (3) 【边框样式】栏。【宽度】下拉列表中选择线条粗细；【样式】下拉列表中选择线型，有虚线、点画线等；【线端】下拉列表中选择虚线类线型的每段线两端的形状，有粗端、圆形、正方形。

2. 修饰文字

图形中的文字包括文本框中输入的文本、图形标题、子标题、脚注、轴标题、坐标轴数值标签、图例标题等。

- (1) 【文本样式】选项卡。
  - ① 【以首选大小预览】。以首选大小显示当前文字式样，选择了字体、字号、颜色后，单击【应用】按钮后，该栏内可以看到实际的文字效果。
  - ② 【字体】栏。在【系列】下拉列表中选择字体；在【样式】下拉列表中选择是否加粗、倾斜或同时加粗、倾斜等；在【大小】下拉列表中选择字号；在【首选大小】下拉列表中选择

选择首选字号；【最小大小】下拉列表中选择最小字号。如果图形不是太小，【大小】中选择【自动】，显示的字号与整个图成比例，首先尝试使用首选字号，再小也不会小于最小字号。如果要插入的图形需要缩小，最好选择较大的字号并加粗，这样打印后的效果较好。

- ③ 【颜色】栏。选择字的颜色，显示在【文本颜色】旁的方块中。
- (2) 【文本布局】选项卡中设置文字布局，即排列方式。可以在预览栏中看到选择结果。
- ① 【对齐方式】栏。在【对齐】下拉列表中选择文字对齐方式；【填充】栏设置文字在它的框架范围中与框架的距离，包括水平和垂直距离。如果对齐方式不是中心对齐，距离设置不当有可能显示不出文字。
- ② 【方向】栏。选择文字排列方向。有【水平】、【从上至下】、【从下至上】、【自定义】4 种方式。

20.2.4 坐标轴的编辑

在【图表编辑器】窗口中，选中坐标轴，打开【属性】窗口。如果没有显示【属性】窗口，按【编辑→属性】顺序单击菜单项打开。坐标轴编辑可能使用到的选项卡有如图 20-17(a) 所示的【刻度】选项卡、如图 20-17(b) 所示的【数字格式】选项卡、如图 20-18(a) 所示的【线】选项卡和如图 20-18(b) 所示的【标签和刻度标记】选项卡。

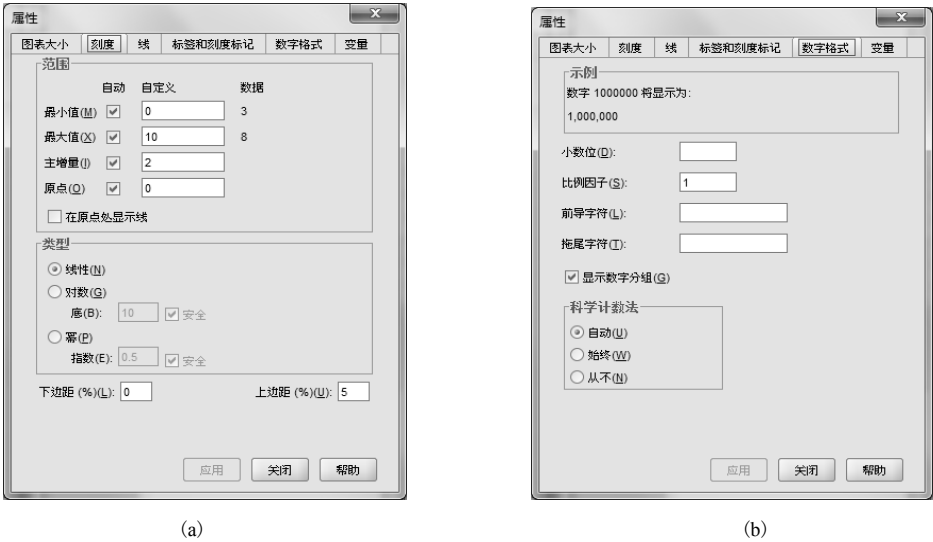


图 20-17 【刻度】和【数字格式】选项卡

1. 【刻度】选项卡

如果选择的坐标轴变量是尺度变量，选择此选项卡。

(1) 【范围】栏。设置坐标轴刻度范围，包括【最大值】、【最小值】、【主增量】(跨度或步长)和【原点】刻度起始位置。

① 选择刻度范围和跨度。它们之间的关系如下：

	自动设置	自定义	数据
【最小值】	系统指定最小值	用户指定最小值	实际数据最小值
【最大值】	系统指定最大值	用户指定最大值	实际数据最大值
【主增量】(跨度)	系统指定主刻度跨度	用户指定主刻度跨度	
【原点】(起始点)	系统指定起始值	用户指定起始值	

② 【在 0 点处显示(直)线】。如果两轴交叉点非 0 点，可以选择此项，在 0 点处显示一条直线。

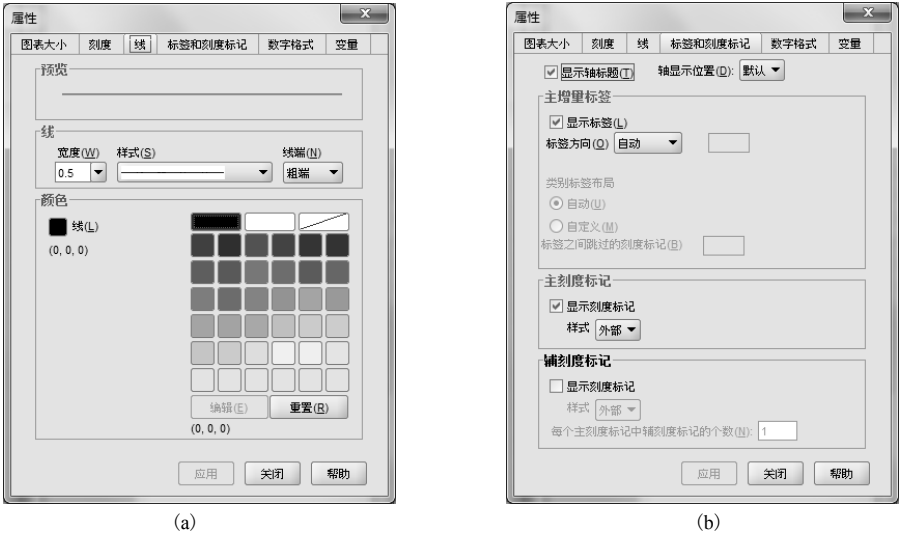


图 20-18 【线】和【标签和刻度标记】选项卡

- (2) 在【类型】栏选择坐标轴变换的方法。为便于得出统计结论，可以对刻度进行转换。
- ① 【线性】。显示线性的未转换的刻度。
  - ② 【对数】。显示对数转换的刻度。在【底】框中输入对数的底，其值必须大于 1。如果选择了【安全】，则不是以  $\log(y)$  作刻度，而是对原刻度的绝对值取对数，再加上符号。
  - ③ 【幂】。显示指数幂刻度。在【指数】框中输入指数，默认值为“0.5”，即开方。如果选择了【安全】则显示安全的刻度，即以原刻度处的数值取指数，而不是对轴变量取指数再作图。
  - ④ 【下边距】、【上边距】。设置图形数据区元素的上、下留白的百分比，默认值为 5%，可以输入的范围为 0~50。如果输入“50”则看不到图形。

2. 【数字格式】选项卡

如果选择的坐标轴变量是尺度变量，选择此选项卡。如果没有出现【属性】窗口，可以从右键菜单中选择。

- ① 【小数位】框。输入刻度标识的小数的位数。
- ② 【比例因子】框。指定比例因子，刻度轴上的每个值除以换算系数。例如，刻度值为 1 000 000、2 000 000、…可以用 1、2、…来代替，同时将“millions”加到轴的标题上。
- ③ 【前导字符】。指定刻度标识的第一个字符，如“\$”。
- ④ 【拖尾字符】。指定刻度标识的最后一个字符，如“%”。
- ⑤ 【显示数字分组】。指定使用千位分节号。
- ⑥ 【科学计数法】栏。有 3 个选项：【自动】、【始终】、【从不】。

3. 【线】选项卡

- (1) 在【线】栏中，在【宽度】下拉列表中选择线的粗细；在【样式】下拉列表中选择线型；在【线端】下拉列表中选择非实线线型的每段线两端的形状。
- (2) 在【颜色】栏中确定坐标轴的颜色。



#### 4. 【标签和刻度标记】选项卡

设置轴上的值标签和刻度线的属性。

(1) 【显示轴标题】。默认选择。

(2) 【轴显示位置】下拉列表。默认为 X 轴在下边、Y 轴在左边。选择【相反】，坐标轴显示到默认位置的对面。

(3) 【主增量标签】栏。

① 【显示标签】。显示坐标轴刻度标签。选中该项，激活【标签方向】下拉列表，可选的排列方向有【自动】、【水平】、【垂直】、【对角线】、【交错排列】，【定制角度】即自定义角度，在后面框中输入旋转的角度。

② 【类别标签布局】。有以下两个选项，只对分类轴有效：

- 【自动】。系统自动放置。

- 【自定义】。【标签之间跳过的刻度标记】指定跳过某些刻度，在后框中输入的数字，决定跳过的标签数。例如，在分类轴上有 A~L 共 12 个分类刻度标签，在框中输入“2”，最后在分类轴上只显示 A、D、G、J 共 4 个标注。

(4) 【主刻度标记】栏。

① 【显示刻度标记】。在【样式】下拉列表中选择显示方式，有 3 项：【外部】，刻度标记在坐标轴外，这是系统默认的标记方式；【内部】，刻度标记在坐标轴内；【穿过】，刻度标记穿越坐标轴线。

② 【辅刻度标记】栏。选择【显示次刻度标记】激活【样式】下拉列表，选择显示方式，各项含义与【主刻度标记】相同。

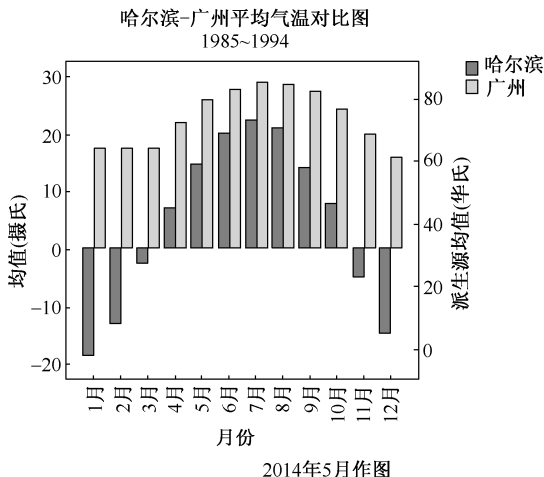
③ 【每个主刻度标记中辅刻度标记的个数】。框中设置每两个主刻度之间的次刻度数。

#### 5. 派生 Y 轴的修饰(派生方法见 20.2.2 节)

选中派生出的 Y 轴，在【属性】窗口中选择【派生轴】选项卡，见图 20-19(a)。在【定义】栏内定义派生轴与尺度轴的对比关系。



(a)



(b)

图 20-19 【派生轴】选项卡及编辑结果

(1) 在【刻度轴】、【派生轴】下面定义原 Y 轴单位与派生 Y 轴单位的比率。例如，【比值】为“1000”，【单位等于】为“1832”意为原 Y 轴 1000 个单位相当于派生 Y 轴 1832 个单位，两个数字框中必须输入正整数。此外，在指定了比值之后，还应该考虑增量的大小。

(2) 【匹配】、【值等于】原 Y 轴上某数值与派生 Y 轴上某数值相对应。例如，【匹配值】框中输入“0”，【值等于】框中输入“32”，即源刻度轴 0 值刻度与派生轴 32 相对应，这种对应关系要根据作图需要。在该选项卡上输入的数据及其效果见图 20-19。

图 20-19(b)所示为哈尔滨-广州月平均气温条形图，原 Y 轴是摄氏度，要在派生轴显示华氏度刻度。X 轴是分类轴刻度标签按自定义排序；主刻度向外，隔 1 个值显示一个，显示轴标题，横轴两端各留 10% 的空间；Y 轴的原点(0℃)显示一直线。

派生轴的设置见图 20-19(a)，效果见图 20-19(b)。

20.2.5 图条的修饰

当生成的图形为条、箱线、误差条、垂线、极差和高低图时，可以对图条进行修饰。图条的修饰在两个选项卡中进行：【条形图选项】选项卡和【深度和角度】选项卡，分别见图 20-20(a)、(b)。以条形图为例说明对类似的条线修饰的方法。

选中图形中某个图例，打开【属性】窗口，选择【条形图选项】选项卡，见图 20-20(a)。

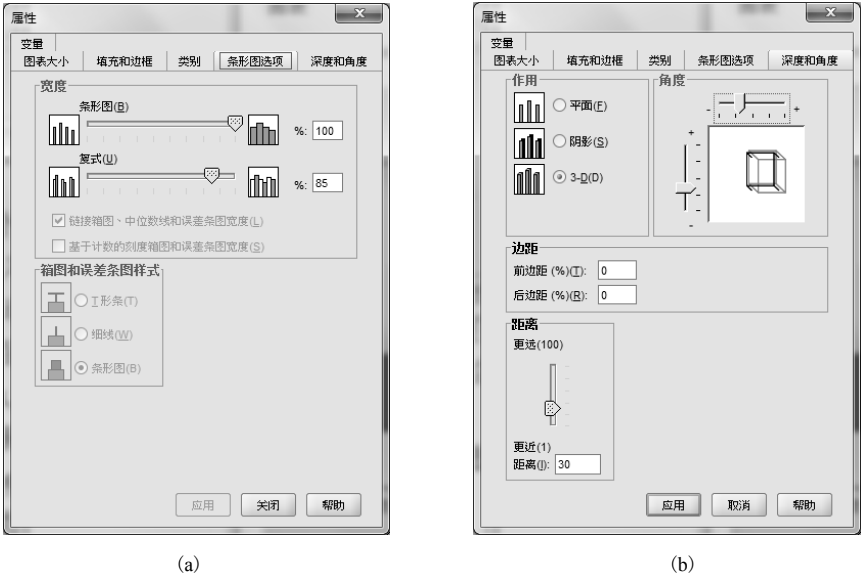


图 20-20 【条形图选项】选项卡、【深度和角度】选项卡

1. 【条形图选项】选项卡

(1) 宽度栏。移动游标或在后框中输入图条的宽度占系统给出范围的百分比。【条形图】调整条图组内间距百分比，【复式】调整条图组间间距百分比。

(2) 【链接箱图、中位数线和误差条图宽度】。在箱图中选择箱体、中线或在误差条图中的误差条时激活此项，移动游标可调整这些元素的宽度。

(3) 【基于计数的刻度箱图和误差条图宽度】。在选择箱图和误差条图后，选择此项，根据分类变量各分类中所含观测量的多少决定每个图条的宽窄。

(4) 【箱图和误差条图样式】栏。选择箱图和误差条图的外伸线的样式。

## 2. 平面效果和立体效果转换

选中某个图例，在【属性】窗口选择【深度和角度】选项卡，见图 20-20 (b)。

(1) 【作用】栏。选择图形效果，【平面】作二维平面条形图；【阴影】作二维但有阴影的条形图；【3-D】作立体图。

(2) 【角度】栏。拖动标尺，选择阴影图和立体图的水平和垂直角度。通过调整角度，立体图表现出不同的深度，调整到预览框中显示的图满意为止。

(3) 【边距】栏。【前边距(%)】和【后边距(%)】框，分别设置立体图前、后两侧留白空间占内框的百分比。

(4) 【距离】栏。改变立体图形视觉上的远近，直观形成图形大小。可以拖动滑块，也可以在【距离】框中输入距离数值，确定图形的大小。距离数值越大，图形越小，看上去就越远。

图 20-21 (a) 所示是阴影条形图，图 20-21 (b) 所示是 3-D 条形图。

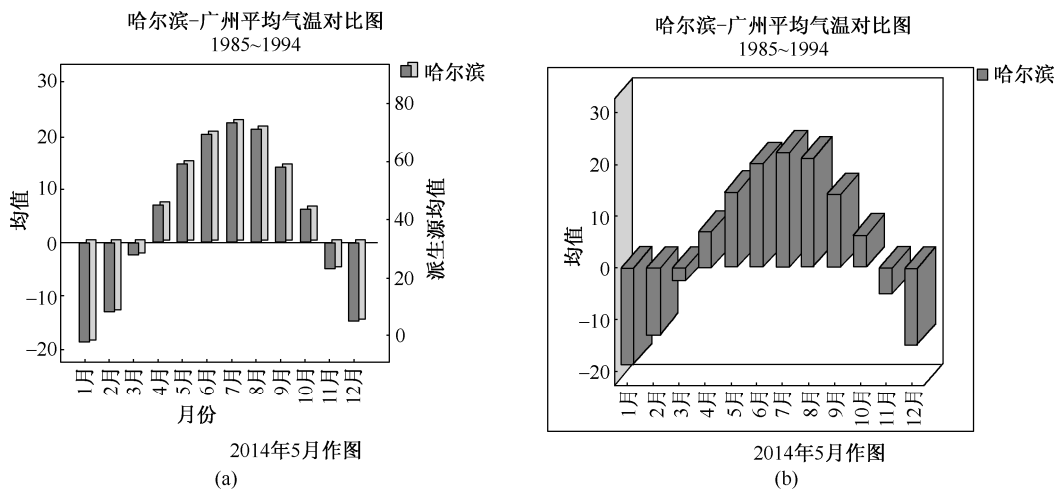


图 20-21 阴影条形图与 3-D 条形图

## 20.2.6 图线的编辑

修饰图线使用【属性】窗口的【内插线】和【线选项】选项卡。选定图线，打开【属性】窗口。当选择了各种图中的线时会自动打开包括这个选项卡的【属性】窗口。

### 1. 【内插线】选项卡

见图 20-22 (a)，确定内插线连线方式。

① 【直线】。系统默认在线图上连接各点的方式为折线，生成的线图就是相邻两点之间用直线线段连接形成的。此为默认方式。

② 【步长】阶梯线。图中水平线穿过每个数值点，垂线连接相同分类值的两个变量数值点。选择左步长、中心步长或右步长来指示数据值在水平线上的位置。

③ 【跳跃】跳跃线。画一条通过每个数据值点的水平线。选择左跳跃、中心跳跃或右跳跃来指示数据值点在水平线上的位置。

【步长】和【跳跃】的画线方法相同，只是【跳跃】没有垂直连接线。

④ 【样条】曲线。用三次方曲线光滑连接相邻的数据值点。

连线还可以应用在散点图、高低图、误差条图等图形上。



图 20-22 【内插线】、【线选项】和【线】选项卡

(2) 【通过缺失值内插线】。连线通过缺失值。

2. 【线选项】选项卡

见图 20-22 (b)，例图见图 20-23。

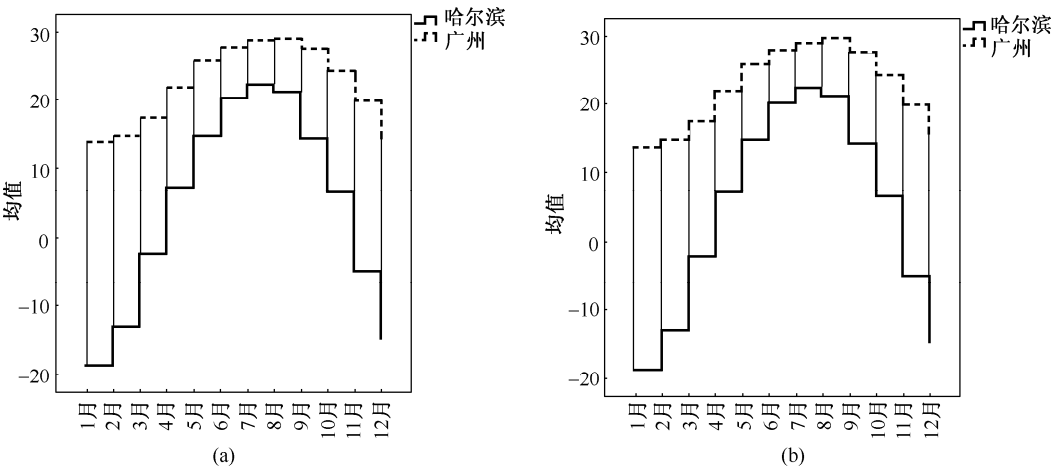


图 20-23 加垂线与突出线的编辑效果

(1) 【显示类别范围条】。用垂线连接同一分类中不同变量的数值。强调同一分类不同变量值间的差异，见图 20-23 (a)，垂线长度反映了两个城市同一月份的温度均值之差的大小，是带有垂线的步长线图，选择的是左步长，点在步长线的左边。

(2) 【投影】(应译为“突出”)栏。选中【显示投影线】，投影线的作用在于从视觉上区分分类轴上的某值两侧曲线。例如：

① 在【起始】下的【类别】下拉列表中选择突出线起始点，突出线将从这个变量值开始。图 20-22 (b)中设置从“6 月”开始。图 20-23 (b)中的加粗线即是突出的部分。

② 【方向】下拉列表。选择突出线的方向，即选择分类变量值向前或向后突出。图 20-23 (b)所示是选择从变量值“6 月”开始，向后加粗的突出线。

### 3. 【线】选项卡

用于编辑所选择的线的宽度、样式、线端样式和颜色。在预览栏中可以看到效果。

## 20.2.7 饼图编辑

为说明饼图编辑的各项功能，以世界各国饮料产量为例作饼图。步骤如下：

(1) 打开数据文件 data20-04。

(2) 按【图形→旧对话框→饼图】顺序单击菜单项，打开【饼图】对话框，选择【个案组摘要】项，单击【定义】按钮，打开【定义饼图：个案组摘要】对话框。

选择碳酸盐和浓缩饮料变量作为饼图的分区表征变量。以变量和作为表现的统计量。用“洲”作为定义分区的变量。单击【标题】按钮，输入两行标题。单击【确定】按钮，作图结果见图 20-25(a)。

### 1. 平面饼图和立体饼图

选中饼图，打开【属性】对话框，选择【深度和角度】选项卡，见图 20-24(a)。

(1) 【作用】栏。【平面】生成的图是平面效果，是系统默认的，还有【阴影】和【3-D】三维效果。

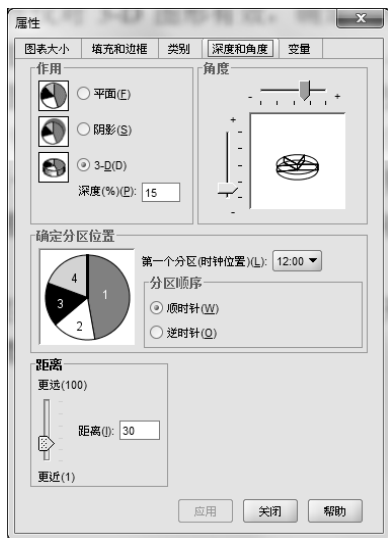
(2) 【角度】栏。对 3-D 效果，仅可移动纵向游标改变纵向观察角度；对阴影效果，可以移动纵向和横向游标改变光源方向。根据预览结果确定想要的最佳角度。

(3) 【距离】栏仅对 3-D 图形有效，确定图形的大小，代表距离远近。选择【3-D】选项后通过输入深度数值，可改变 3-D 饼图的高度。

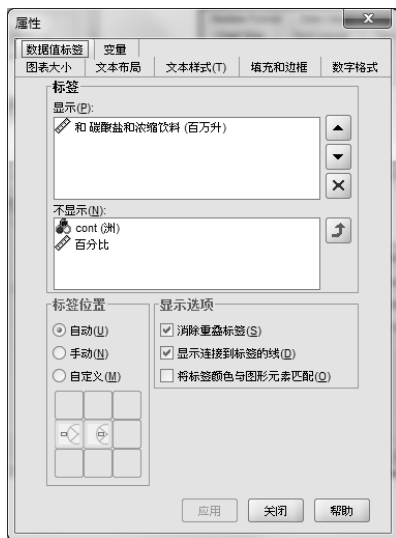
(4) 【确定分区位置】栏。确定扇形位置和排列方向。

① 【第一分区(时钟位置)】。以钟表盘的方式确定第一个扇面的位置。在后面的下拉列表中选择第一分区的时间。如果选择“12:00”，则从时钟 12 点位置开始排列扇形。

② 【分区顺序】栏。选择饼图中扇形的排列方向，有【顺时针】或【逆时针】排列。



(a)



(b)

图 20-24 饼图的【深度和角度】和【数据值标签】选项卡

2. 数值标签

选择饼图，在右键菜单中选择【数据值标签】，打开【属性】窗口【数据值标签】选项卡，见图 20-24(b)。

(1) 【标签】栏。决定显示什么内容的标签。【显示】框中是已经显示的标签。【不显示】框中是没有显示在图中的，选中变量拖入【显示】框的均可以在饼图中显示。在该选项卡中可以将【不显示】框中的“百分比”、“洲”的值也显示在饼图的各扇形区中。方法是选择其中一个，单击向上箭头按钮，将想显示的项移到【显示】框中。单击【确定】按钮，实现设置。例如，想显示百分比，不显示具体数值，选择碳酸饮料和浓缩饮料(百万升)，单击红色叉子按钮，将其移到【不显示】框内，将“百分比”通过向上箭头移到【显示】框内。单击【应用】按钮。得到图 20-25(b)。

(2) 【标签位置】栏。用户可以自行指定标签的位置。【自动】由系统决定，生成标签的位置；选择【手动】，可以在饼图中用鼠标拖拽标签到任何想要的位置；选择【自定义】，可以在下面的九格表中选择一种列在其中的显示位置。

(3) 【显示选项】栏。选择标签显示方式，有 3 个选项：【消除重叠标签】、【显示连接到标签的线】、【将标签颜色与图形元素匹配】。这 3 个选项很容易理解，不再赘述。

3. 分离扇面的饼图

选择一个扇面，在右键菜单中选择【分解分区】，被选择的扇面脱离饼图单独显示。

图 20-25(c)所示是将“西欧”的扇形分离出来的饼图。在【数据值标签】选项卡中，标签位置的设置方式为【手动】，单击【应用】按钮后，手动拖拽数值标签到想要的位置。

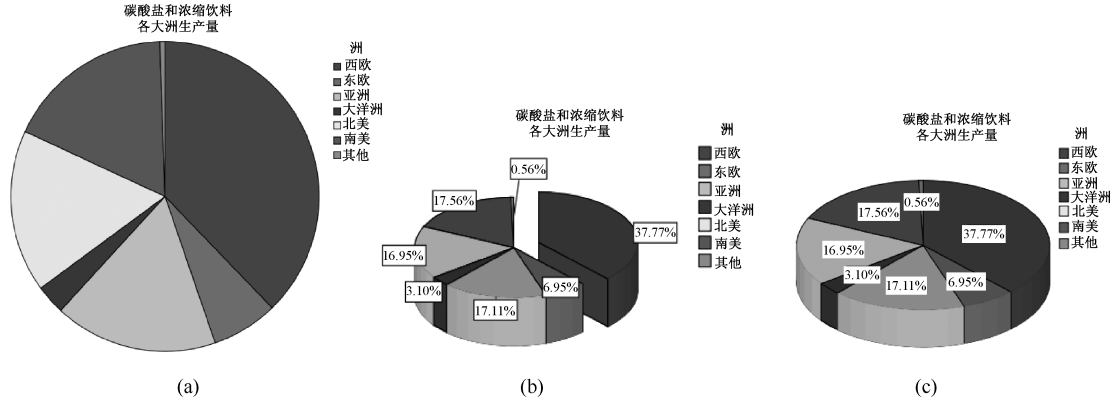


图 20-25 饼图的编辑效果

20.2.8 散点图的编辑

本小节作最简单的散点图，用以说明编辑散点图的各种工具及其功能和方法。

数据文件 data20-02 中是 451 名青少年体质测量数据，其中有“肺活量”和“体重”两个变量。为了探讨这两个变量之间的关系，作散点图继续初步观察。

打开数据文件 data20-02，按【图形→旧对话框→散点/点状图】顺序单击菜单项，打开【散点/点状图】对话框。

(1) 选择简单分布。

单击【定义】按钮，打开【简单散点图】对话框。选择“体重”变量送入【Y 轴】框，选

择肺活量变量送入【X 轴】框。单击【标题】按钮，打开【标题】对话框，输入图形标题。

单击【确定】按钮，在输出窗中生成图形，见图 20-26(a)。

(2) 选择重叠分布。

单击【定义】按钮，打开【重叠散点图】对话框。第一组送“体重”变量作 Y 轴变量，送“肺活量”作 X 轴变量；第二组，送“身高”变量作 Y 轴变量，送“肺活量”变量作 X 轴变量。

单击【标题】按钮，打开【标题】对话框，输入图形标题。

单击【确定】按钮，在输出窗中，生成图形，见图 20-26(b)。

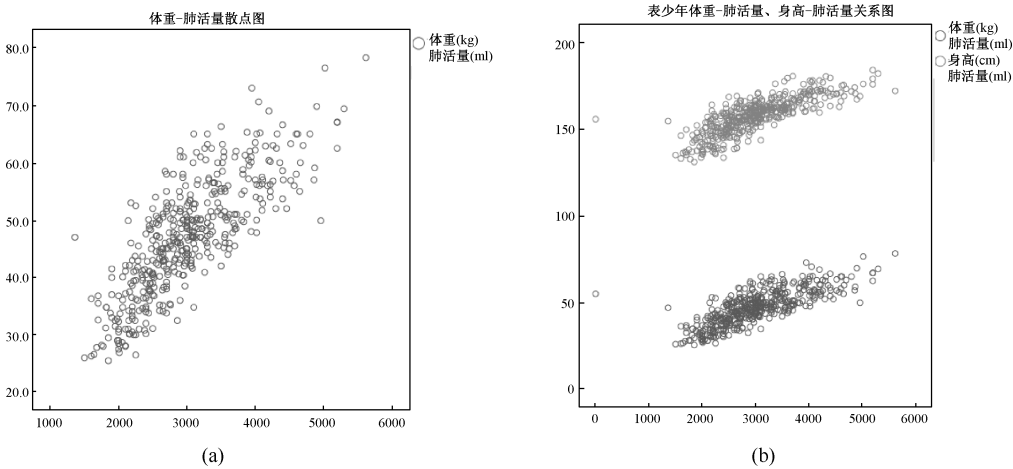


图 20-26 简单散点图和重叠散点图

常用的编辑功能如下：

(1) 点样式的编辑。各种类型散点图的【属性】窗口都有【标记】选项卡，见图 20.27(a)，编辑方法都相同。

选中图中的点，打开【属性】窗口，选择【标记】选项卡。可以选择点的【类型】、【大小】、【边框宽度】以及【颜色】等。

(2) 散点图类型在【属性】窗口【变量】选项卡中进行。在【元素类型】下拉列表中选择可以转化成的其他类型的散点图和其他图形。

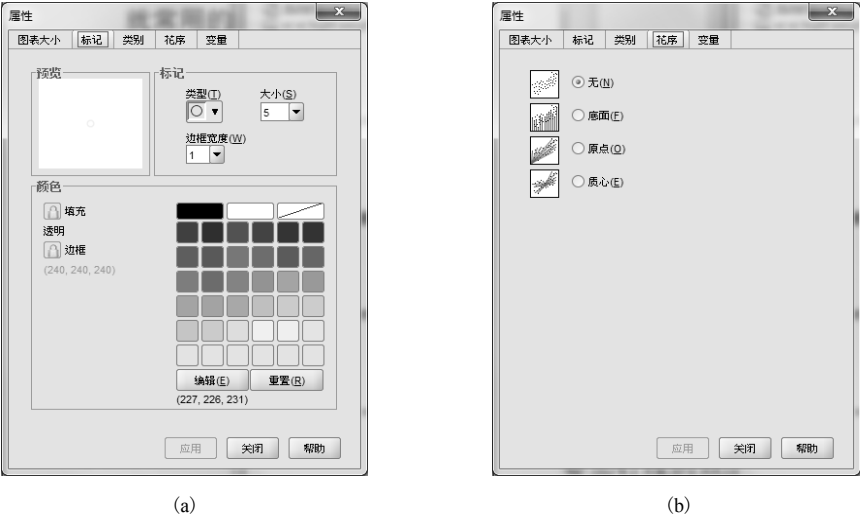


图 20-27 【标记】与【花序】选项卡

3. 花序

钉线是指从每个数据点到所选定位置画的线段，它可用来观察数据点的差异。在图形框中，双击要加钉线的点，打开【属性】窗口，选择【花序】选项卡，见图 20-27(b)。

- ① 【无】。系统默认无钉线，即生成的原图形，见图 20-26 的(a)、(b)。
- ② 【底面】。对平面散点图，钉线为每个数据点到 X 轴的垂直线，对 3-D 散点图，为每个数据点到 XZ 轴平面的连线。
- ③ 【原点】。从每个数据点到原点的连线。
- ④ 【质心】。从每个数据点到全部数据质心的连线，质心的坐标是 XY(Z)轴上两三个变量值的加权平均数，若其中任一变量中有缺失值，要从计算中剔除。改变轴的刻度不影响质心点的计算。

图 20-28 所示是使用数据文件 data20-02 所作的简单散点图、重叠散点图的钉线 3 种方式的效果，容易看出 3 个选项的含义。图 20-28(a)所示是底面钉线效果，图 20-28(b)所示是散点图的原点钉线的效果，图 20-28(c)所示是质心散点图的效果。

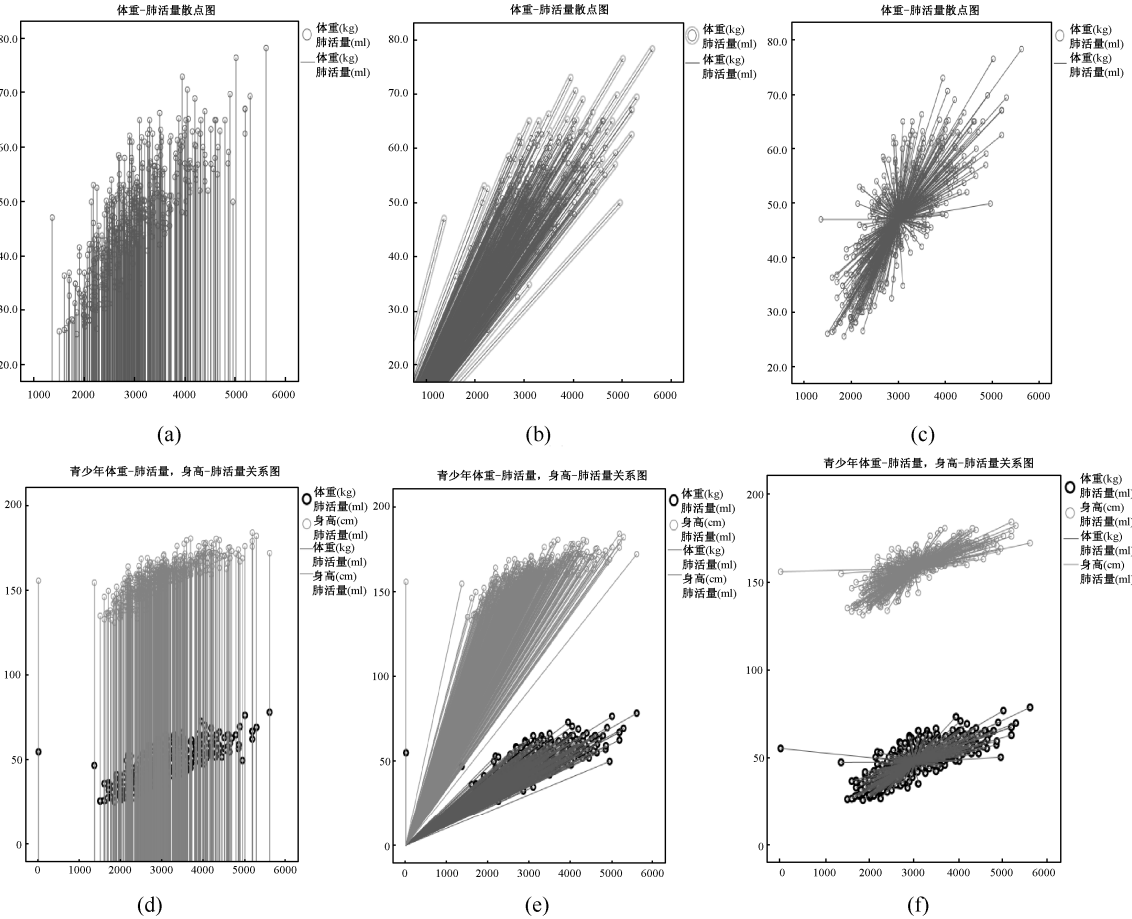



图 20-28 简单与重叠散点图的底面、原点、质心钉线花絮的效果

4. 拟合线的生成与修饰

(1) 生成拟合线。在【图表编辑器】窗口，选择【散点图】，在工具栏或右键菜单中选择



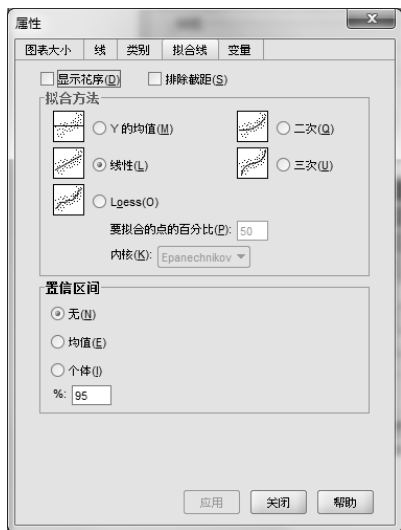
对所有点产生拟合直线；选择  分类产生拟合直线。图例下方显示线性拟合统计量  $R$ ，该值越大拟合得越好。

使用数据文件 data20-02 作出的男、女肺活量(X 轴)和体重(Y 轴)散点图和拟合直线的结果，见图 20-30(a)。男性直线拟合优度  $R^2=0.750$ ，女性直线拟合优度  $R^2=0.476$ 。

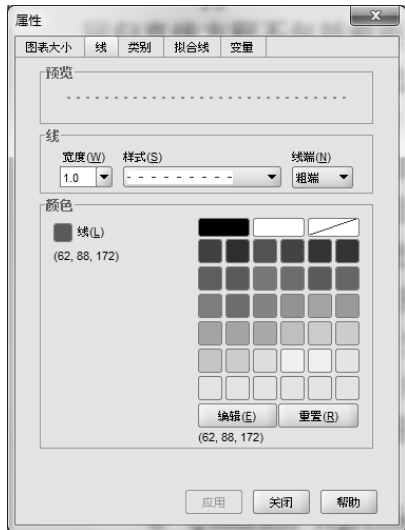
如果在【拟合线】选项卡中选择了【二次】，拟合结果见图 20-30(b)。

男性二次曲线拟合优度  $R^2=0.755$ ，女性二次曲线拟合优度  $R^2=0.498$ 。

(2) 编辑拟合线。在散点图中选中拟合线，打开【属性】窗口选择【线】选项卡，见图 20-29(b)。



(a)



(b)

图 20-29 【拟合线】选项卡和【线】选项卡

### ① 两个复选项：

- 【显示花序】。显示拟合线到每个点的垂直连线。
- 【排除截距】。修改拟合线，使之通过原点，即回归直线方程不包括截距。

② 【拟合方法】栏可以选择 5 种拟合方式：【Y 的均值】、【线性】、【Loess】对所有数据继续拟合，【二次】、【三次】即拟合二次或三次曲线。从中选择一种方法，单击【应用】按钮，输出窗中显示拟合结果。每选择一次，拟合一次。可以比较几次不同拟合的  $R^2$  值，选择其值最大的拟合结果，即最佳拟合效果。如果已知数据趋于线性回归直线、二次回归曲线和三次回归曲线，则可以直接从下面的选项中拟合数据；如果不了解数据集趋于何种曲线，可以从直线开始一个个试拟合。图 20-30 所示是直线拟合和二次曲线拟合的结果。可以拟合的曲线选项有：

- 【Y 的均值】。生成一条 Y 轴数值的平均线。
- 【线性】。线性回归直线。根据最小平方方法，用线性回归直线对散点图中的数据点进行最佳拟合。这是系统默认的方式。
- 【二次】。根据最小平方方法，用二次回归曲线拟合散点图中的点。
- 【三次】。根据最小平方方法，用三次回归曲线拟合散点图中的点。
- 【Loess】。局部加权回归散点修匀法。用迭代加权最小平方方法拟合，至少需要 13 个点。

- **【要拟合点的百分比】**。指定用于拟合的数据占总数的百分比，默认值为“50%”。
- **【内核】**。在其后的下拉列表中选择所需要的核函数。

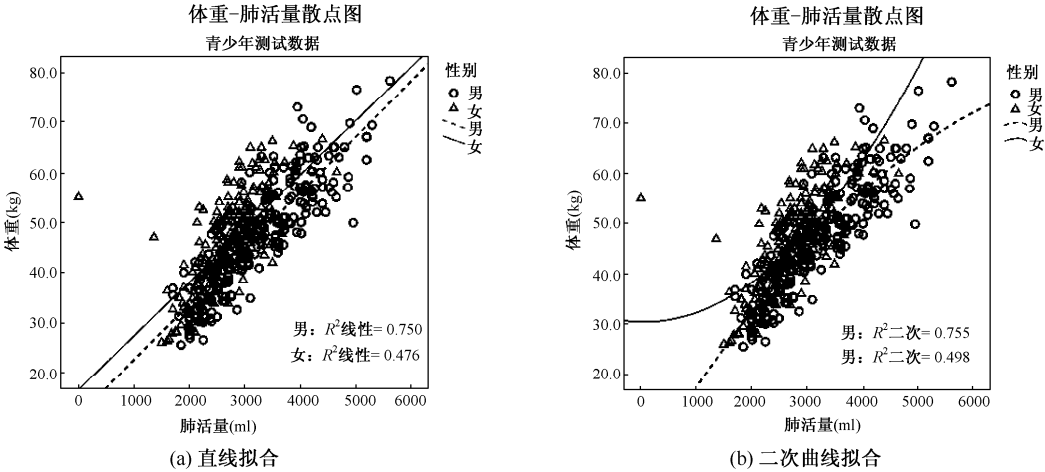


图 20-30 直线拟合与二次曲线拟合与修饰的结果

图 20-30(a)、(b)的  $R^2$  值比较结果：直线拟合的两个  $R^2$  分别为 0.75、0.476，二次曲线拟合的  $R^2$  分别为 0.755 和 0.498。虽然看起来二次拟合比较好，但是差别不大，还应该取更多的观测进行拟合。本例只取了 451 个观测的结果。

**注意：**缺失值会对图形有较明显的影响，作图前一定要定义好缺失值，使之不参与绘制图形。

③ **【置信区间】** 栏。选定拟合线的可信区间。**【无】** 不生成可信区间线；**【均值】** 生成平均值的可信区间线；**【个体】** 生成单个观测量的可信区间线。**【均值】** 和 **【个体】** 选项需要指定可信区间百分比数，默认值为“95%”。

20.2.9 文件管理

1. 保存图形模板

用户将生成和完成编辑的图形保存为模板文件，以后在生成新的图形时，可调用模板文件，新生成图形的格式与模板中的图形格式一致，省去了许多烦琐的编排工作。

在**【图表编辑器】**窗口中按**【文件→保存图表模板】**顺序单击菜单项，打开**【保存图表模板】**对话框，见图 20-31。可以选择想保存在模板文件中的图形要素。当前图形所有的图形要素都以折叠菜单形式显示在对话框中。可选择保存的种类和细项大致有以下内容：

(1) **【所有设置】**。保存所有图形元素的设置

① **【布局】**。确定了模板图形的版面编排，包括图形大小、图形的宽高比例、图形各个边框的大小位置和显示顺序、图形的方向等。



图 21-31 **【保存图表模板】**对话框

② **【文本内容】**。包括图形标题、轴标题、注脚、注释等元素。不包括数据值标签的文字。

③ **【样式】**。包括文字格式；非图形元素样式，如填充和边界样式、线型、条、点的样式、是否有坐标线、背景板等；图形元素样式，包括简单无分组图形元素样式；线条和标记的样式；条形图宽度、一类若干条；箱图、误差条图中的线、细线或条的样式；所占宽度针对不同的图形有不同的选择项。另外，还可以对 3-D 图的深度、距离、旋转等进行设置，及饼图第一分区的定义、起始位置等，还有点图的花絮设置等。

④ **【轴】**栏。可选择的有：**【刻度轴】**的范围，刻度轴上、下空白空间，刻度轴的数值类型，主刻度标记和每个主刻度之间次刻度的数量，显示派生轴；**【刻度设置】**起始位置，刻度轴的日期或时间的版式，刻度轴的数值版式；**【分类轴】**左、右的空白空间，合并分类的设置；分类值标签的显示方式，分类标签的排列方式等。

⑤ **【统计量】**栏。可选择显示拟合线、参照线、连线和直方图正态曲线等。

(2) 单击**【全部展开】**按钮，所有折叠都打开，显示项目框中的所有项目。

(3) **【模板说明】**栏。输入对本模板的描述文字。

(4) 单击**【重置】**按钮，恢复到选择前的状态，出现选择要保存的项目。选择完成后单击**【继续】**按钮打开**【保存】**对话框，选择保存位置和文件名，单击**【保存】**按钮，保存完成。

## 2. 应用图形模板

在**【图表编辑器】**窗口中，套用某个图形模板，按**【文件→应用图形模板】**顺序单击菜单项，打开**【应用图形模板】**对话框，选择需要的图形模板。

## 习 题 20

1. 各种图形(条形图、饼图、散点图等)是由哪些成分组成的？

2. 数据文件 data20-03 中是世界各地气候、人口状况数据，作出宗教信仰饼图并修饰。对世界上不同宗教所占百分比饼图，进行以下调整：显示每个扇面的文字、数值和百分比；合并小于 5% 的扇面。

# 附录 A 标准化、距离和相似性的计算

SPSS 的许多过程中都使用了距离和相似性、不相似性的计算，例如聚类分析、尺度分析等都会用到。

## 1. 对于等间隔测度的变量(尺度变量，测度类型为 Scale)计算距离的方法

约定：距离或相似性的公式中  $x$ 、 $y$  均表示  $n$  维空间中的两个点， $x_i$  是  $x$  点的第  $i$  个变量的值， $y_i$  是  $y$  点第  $i$  个变量的值。

(1) Euclidean distance(欧氏距离)。两项之间的差是每个变量值之差的平方和之平方根。

$$EUCLID(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

(2) Squared Euclidean distance(欧氏距离平方)。两项间的距离是每个变量值之差的平方和。

$$SEUCLID(x, y) = \sum_i (x_i - y_i)^2$$

(3) Cosine(cos 相似性测度)。计算值向量间的余弦，值范围是 $-1 \sim 1$ ，用 0 值表明两向量正交(相互垂直)。

$$COSINE(x, y) = \frac{\sum_i (x_i y_i)^2}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

(4) Pearson correlation(皮尔逊相关)。计算值向量间的相关，Pearson 相关是线性关系的测度，范围是 $-1 \sim 1$ 。0 值表明没有线性关系。

$$CORRELATION(x, y) = \frac{\sum_i (Z_{x_i} Z_{y_i})^2}{n - 1}$$

(5) Chebychev(切贝谢夫距离)。两项间的距离用最大的变量值之差的绝对值表示。

$$CHEBYCHEV(x, y) = \text{Max}_i |x_i - y_i|$$

(6) Block(布洛克距离)。两项之间的距离是每个变量值之差的绝对值总和。

$$BLOCK(x, y) = \sum_i |x_i - y_i|$$

(7) Minkowski(明可斯基距离)。两项之间的距离是各变量值之差的  $p$  次方幂的绝对值之和的  $p$  次方根。

$$MINKOWSKI(x, y) = \sqrt[p]{\sum_i |x_i - y_i|^p}$$

(8) Customized(自定义距离)。两项之间的距离用各项值之间差值绝对值的  $p$  次幂之和的  $r$  次方根表示。 $p, r$  可以自己指定。

$$MINKOSKI(x,y)=\sqrt[r]{\sum_i |x_i - y_i|^p}$$

2. 两个计数变量的不相似性测度的方法

(1) Chi-square measure( $\chi^2$  测度)。用卡方值测度不相似性。该测度是假设两个集的频数相等进行的卡方检验，测度产生的值是卡方值的平方根。这是系统默认的对计数变量的不相似性测度方法，是根据被计算的两个观测量或两个变量总频数计算其不相似性。期望值来自观测量或变量  $(x, y)$  的独立模型，其中  $E(x_i)$  和  $E(y_i)$  为频数期望值。

$$CHISQ(x,y)=\sqrt{\frac{\sum_i \frac{(x_i - E(x_i))^2}{E(x_i)}}{E(x_i)} + \frac{\sum_i \frac{(y_i - E(y_i))^2}{E(y_i)}}{E(y_i)}}$$

(2) Phi-square measure(两组频数间的  $\Phi^2$  测度)。该测度考虑了减少样本量对测度值的实际预测频率减少的影响。该测度把  $\Phi$  平方除以联合频数的平方根，使不相似性的卡方测度规范化。该值与参与计算不相似性的两个观测量，或两个变量的总频数无关。

$$PH2(x,y)=\sqrt{\frac{\sum_i \frac{(x_i - E(x_i))^2}{E(x_i)}}{E(x_i)} + \frac{\sum_i \frac{(y_i - E(y_i))^2}{E(y_i)}}{E(y_i)}} \cdot \frac{1}{\sqrt{N}}$$

3. 二值变量的距离或不相似性测度的约定

(1) 首先应该明确，对二值变量，系统默认用 1 表示某特性的出现(或发生、存在等)，用 0 表示某特性不出现(或不发生、不存在)。

表 A-1 四格表

第一特性	第二特性	
	发生	不发生
发生	$a$	$b$
不发生	$c$	$d$

(2) 对二值变量的相似性或不相似性测度都基于一个四格表，见表 A-1。

如果对观测量进行计算，则对所有“变量对”做一遍四格表。如果对变量进行计算，则对所有“观测量对”做一遍四格表。对每个四格表按所选择的方法进行一次距离参数的计算，这样形成距离矩阵。例如，分析变量  $V, W, X, Y, Z$ ，观测量 1 的 5 个变量值顺序为 0、1、1、0、1；观测量 2 的 5 个变量值顺序为 0、1、1、0、0，如下面的表 A-2 所示。

表 A-2 例题数据中的两个观测量及对应的四格表

分析变量 观测量号	$V$	$W$	$X$	$Y$	$Z$
	0	1	1	0	1
2	0	1	1	0	0

		第二特性	
		发生	不发生
第一特性	发生	$a=2$	$b=1$
	不发生	$c=0$	$d=2$

两个事件都发生的有两个变量  $W, X$ ，相应的四格表的  $a$  为 2；两个事件都不发生的有变

量  $V$ 、 $Y$ ，因此  $d=2$ ；事件 1 发生，事件 2 不发生的是变量  $Z$ ，因此  $b=1$ ；事件 1 不发生，事件 2 发生的没有，因此， $c=0$ 。相应的四格表读者自己可以做出。

(3) 对二值数据的相似性或不相似性测度，或二值变量距离测度算法的分类：

- ① 匹配系数的计算，包括：RR、SM、SS1、RT、JACCARD、DICE、SS2、K1、SS3。
- ② 与条件概率有关的测度包括：K2、SS4、HAMANN。
- ③ 与预测特性有关的测度包括：Y、Q、LAMBDA、D。
- ④ 其他距离、相关等测度包括：BEUCLID、BSEUCLID、SIZE、PATTERN、BSHAPE、

OCHIAI、SS5、PHI 等。

(4) 在下面给出的公式中，作为自变量的  $x$ ， $y$  不一定指两个变量。因为用观测量之间的相似性或距离可以进行观测量聚类，那么  $x$ 、 $y$  就指两个观测量；作变量聚类，要计算两个变量的距离或相似性、二值相关，那么这里的  $x$ ， $y$  就是两个变量。

(5) 另外，表中联合发生的指  $a$ ，联合不发生的指  $d$ ，所有匹配的指  $a+d$ ，所有不匹配的指  $c+b$ ， $n=a+b+c+d$ 。表 A-3 说明匹配系数。

(6) 按权重和分子分母特征归纳各计算方法如表 A-3。

4. 二值变量的距离或不相似性测度的方法

(1) Euclidean distance，二值欧氏距离。根据四格表计算  $\text{SQRT}(b+c)$ 。 $b$  和  $c$  表示事件在 一项中发生，在另一项不发生的对角单元，其最小值为 0，无上限。

(2) Squared Euclidean distance，二值欧氏距离平方。计算的是不匹配事件的数目，其最小值为 0，无上限，数值等于  $b+c$ 。

(3) Size difference，不对称指数，其值范围在 0~1 之间。

$$SIZE(x,y)=\frac{(b-c)^2}{n^2}$$

表 A-3 匹配系数计算中的权重关系及分子、分母特征表

		分子中不包括联合不发生的 $d$	分子中包括联合不发生的 $d$
All matches included in denominator 分母中包括所有匹配的 ( $a$ 、 $d$ )	给匹配与不匹配的权重相等	RR	SM
	给匹配的双倍权重		SS1
	给不匹配的双倍权重		RT
Joint absences excluded from denominator 联合不发生的的不包括在分母中 ( $d$ )	给匹配与不匹配的权重相等	JACCARD	
	给匹配的双倍权重	DICE	
	给不匹配的双倍权重	SS2	
All matches excluded from denominator 分母中剔除所有匹配的 ( $a$ 、 $d$ )	给匹配与不匹配的权重相等	K1	SS3

(4) Pattern difference，不相似性测度。范围为 0~1。根据四格表计算  $bc/n^2$ ，其中  $b$  和  $c$  表示事件在 一项中发生，在另一项中不发生的对角单元。 $N$  是观测量或变量总数。

(5) Variance，方差不相似性测度。根据四格表计算  $(b+c)/4n$ ，Variance 值范围为 0~1。

(6) Dispersion，是一个相似性指数。其范围为-1~1。

$$DISPER(x,y)=\frac{ad-bc}{n^2}$$

(7) Shape，距离测度。范围无上下限。

$$BSHAPE(x, y) = \frac{n(b+c) - (b-c)^2}{n^2}$$

(8) Simple matching, 匹配数与值的总数的比值。它给匹配与不匹配以相同的权重。

$$SM(x, y) = \frac{a+d}{n}$$

(9) Phi 4-point correlation, 皮尔逊相关系数二值模拟, 其值范围为-1~1。

$$PHI(x, y) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

(10) Lambda 数, 是 Goodman and Kruskal 的  $\lambda$ , 是一种相似性测度。当预测方向同等重要时该系数估计的是用一项预测另一项的误差降低的比例。其值范围为 0~1。

$$LAMBDA(x, y) = \frac{t_1 - t_2}{2n - t_2}$$

其中  $t_1 = \text{Max}(a, b) + \text{Max}(c, d) + \text{Max}(a, c) + \text{Max}(b, d)$ ,  $t_2 = \text{Max}(a+c, b+d) + \text{Max}(a+d, c+b)$ 。

(11) Anderberg'D 统计量, 类似于  $\lambda$ , 该指数取决于用一项预测另一项(在两个方向上进行预测)的误差降低的实际数值。其值范围为 0~1。

$$D(x, y) = \frac{t_1 - t_2}{2n}$$

其中  $t_1, t_2$  定义与(10)中的定义相同。

(12) Dice, 该指数中剔除了联合不发生, 给匹配以双倍权重。类似 Czekanowski 或 Sorensen 测度。

$$DICE(x, y) = \frac{2a}{2a + b + c}$$

(13) Hamann, 相似性测度。该指数是匹配数减去不匹配数除以总项数。其值范围是-1~1。

$$HAMANN(x, y) = \frac{(a+d) - (b+c)}{n}$$

(14) Jaccard, 是一个不考虑联合缺席( $d$ )的指数。它给匹配与不匹配以相等的权重, 类似相似比。

$$JACCARD(x, y) = \frac{a}{a + b + c}$$

(15) Kulczynski 1, 是联合出现与非匹配数的比。该指数有下界 0, 无上界。理论上对无不匹配的情况( $b=0, c=0$ )没有定义, 然而, 当值是没有定义的或大于 9999.999 时, 软件赋值给该指数一个武断值 9999.999。

$$K1(x, y) = \frac{a}{b+c}$$

(16) Kulczynski 2, 相似性测度。该指数根据某特性在一项中出现的条件概率给出在其他项中出现的概率。计算该指数时, 每一项作为其他项的预测值时, 各值取其平均数。

$$K2(x, y) = \frac{a / (a + b) + a / (a + c)}{2}$$

(17) Lance and Williams, 根据四格表计算  $(b+c)/(2a+b+c)$ , 其中  $a$  表示与事件在两项中都发生相对应的单元,  $b$  和  $c$  表示事件在两项中发生而在另一项中不发生的对角单元。该测度的值的范围为 0~1。有如我们所知的 Bray-Curtis 非度量系数。

(18) Ochiai, 该指数是余弦相似性测度的二元形式。范围为 0~1。

$$OCHIAI(x, y) = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

(19) Rogers and Tanimoto, 是一个给不匹配的  $(b, c)$  双倍权重的指数。

$$RT(x, y) = \frac{a + d}{a + d + 2(b + c)}$$

(20) Russel and Rao, 是内积(点积)的二元形式。对匹配与不匹配都给予相等的权重, 是二元相似数据的系统默认方法。

$$RR(x, y) = \frac{a}{a + b + c + d}$$

(21) Sokal and Sneath 1, 给匹配以双倍权重的一种指数。

$$SS1(x, y) = \frac{2(a + d)}{2(a + d) + b + c}$$

(22) Sokal and Sneath 2, 给不匹配以双倍权重的一种指数, 且不考虑联合缺席的情况。

$$SS2(x, y) = \frac{a}{a + 2(b + c)}$$

(23) Sokal and Sneath 3, 匹配与不匹配的比。该指数下界为 0, 无上界。理论上, 对无不匹配的情况没有定义。当值为未定义或大于 9999.999 时, 软件给予该指数一个特定常数 9999.999。

$$SS3(x, y) = \frac{a + d}{b + c}$$

(24) Sokal and Sneath 4, 同一匹配状态(某特性出现或不出现)在另一项出现或不出现的条件概率。计算该指数时, 每一项作为其他项的预测值时, 各项值取其平均数。该指数范围为 0~1。

$$SS4(x, y) = \frac{a / (a + b) + a / (a + c) + d / (b + d) + d / (c + d)}{4}$$

(25) Sokal and Sneath 5, 该指数是正负匹配的条件概率的几何平均数的平方。它独立于项编码。其值范围为 0~1。

$$SS5(x, y) = \frac{ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

(26) Yule's Y, 该指数是 2×2 表交叉比的函数, 且独立于边际总和, 范围为 -1~1。有如我们所知的综合系数。



$$Y(x, y) = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

(27) Yule's Q, 该指数是 Goodman 和 Kruskal  $\gamma$ (gamma) 的特殊事件, 是 2×2 表交叉比的函数, 且独立于边际总和。其值范围为 -1~1。

$$Q(x, y) = \frac{ad - bc}{ad + bc}$$

## 5. 对数据进行标准化的方法

① Z scores, 把数值标准化到 Z 分数。标准化后变量均值为 0, 标准差为 1。系统将每一个值减去正被标准化的变量或观测量的均值, 再除以其标准差。如果原始数据的标准差为 0, 则将所有值置为 0。

② Range -1 to 1, 把数值标准化到 -1 至 1 范围内。选择该项, 对每个值用正在被标准化的变量或观测量的值的范围去除。如果范围为 0, 所有值不变。

③ Maximum magnitude of 1, 把数值标准化到最大值为 1。该方法是把正在标准化的变量或观测量的值用最大值去除。如果最大值为 0, 则用最小值的绝对值除再加 1。

④ Range 0 to 1, 把数值标准化到 0 至 1 的范围内, 对正在被标准化的变量或观测量的值减去正在被标准化的变量或观测量的最小值, 然后除以范围。如果范围是 0, 将所有变量值或观测值设置为 0.5。

⑤ Mean of 1, 把数值标准化到均值的一个范围内。对正在被标准化的变量或观测量的值除以正在被标准化的变量或观测量的值的均值。如果均值是 0, 对变量或观测量的所有值都加 1, 使其均值为 1。

⑥ Standard deviation of 1, 把数值标准化到单位标准差。该方法对每个值除以正在被标准化的变量或观测量的标准差。如果标准差为 0, 则这些值保持不变。

# 附录 B 数据清单

数据编号	数据名称	出现页码
data02-01.sav	1969—1971 年 美国某银行 474 雇员状况数据	40, 65, 90, 91
data02-02.sav	青少年身高和体重 (1)	53, 54
data02-02a.txt	data02-02.sav 以 ASCII 格式保存 (固定格式, 有列间隔)	54, 55
data02-02b.txt	data02-02.sav 以 ASCII 格式保存 (固定格式, 无列间隔)	54, 55, 56
data02-03.txt	啤酒数据 (固定格式, ASCII 码, 有列间隔)	54, 55, 91
data02-04.txt	啤酒数据 (自由格式, ASCII 码, 有列间隔)	54, 55, 59
data02-05.sav	青少年身高数据 (1)	66
data02-06.sav	青少年身高数据 (2)	68
data02-07.sav	青少年身高和体重 (2)	68
data02-08.sav	青少年身高数据 (3)	70, 71
data02-08a.sav	文件合并结果数据	71
data02-09.sav	青少年体重数据	70, 71
data02-10.sav	237 个人的身高体重数据	75
data02-11.sav	某银行雇员工资和受教育状况 (data02-01.sav 简化版)	76, 77
data02-11a.sav	data02-11.sav 重新编码运行结果	76
data02-12.sav	顾客对 17 种汽车评价	79
data02-12a.sav	data02-12 数据转置结果	79
data02-13-1.sav	学生学号、身高、体重及 3 门课程分数	81
data02-13-1a.sav	data02-13-1.sav 数据结构重建 (ScoreA、ScoreB 和 ScoreC 变量组转换成 一个观测量组, 索引变量为顺序值)	84
data02-13-1b.sav	data02-13-1.sav 数据结构重建 (保留固定变量 <i>h</i> 、 <i>w</i> , 索引变量值为 原始变量名)	84
data02-13-1c.sav	data02-13-1.sav 数据结构重建 (不保留固定变量, 索引变量值为原 始变量名)	84
data02-13-2.sav	学生身高、体重及文理科分数数据	85
data02-13-2a.sav	data02-13-2.sav 数据结构重建 (以文科和理科成绩各自生成变量) 转换后的数据	85
data02-14.sav	两班学生 3 门课程成绩	86
data02-14a.sav	data02-14.sav 数据结构重建	86
data02-15.sav	期中和期末的 3 门课程成绩	86
data02-16.sav	3 个市场、7 种商品价格	91
data02-17.xls	中国女排档案	62
data04-01.sav	随机变量分布函数例 1 题结果	119
data04-02.sav	随机变量分布函数例 2 题结果	120
data05-01.sav	生日日期型数据	127
data05-01a.sav	data05-01 中变量类型转换结果	127
data05-02.sav	数值型数据及转换成日期型数据的结果	128
data05-02a.sav	data05-02 中变量类型转换结果	128
data05-03.sav	字符型数据	129
data05-03a.sav	data05-03 字符型转换成日期型数据结果	130
data05-04.sav	生日数据	131

数据编号	数据名称	出现页码
data05-04a.sav	字符型日期转换成数值型结果	131, 132
data05-04b.sav	生日计算年龄的结果	132
data05-05.sav	生日数据	132
data05-05s.sav	从生日提取月份的结果	133
data06-01.sav	不同性别、年龄、婚姻状况的生活方式和首选早餐的调查	144
data06-02.sav	公务员晚饭后的活动内容	149, 153, 157
data06-03.sav	公务员晚饭后的三个主要活动	150, 158
data06-04.sav	不同受教育年限的各种职务的平均初始工资	160
data07-01.sav	1991 年美国社会调查	163, 164, 179
data07-02.sav	1985 年美国 50 个州犯罪记录	167
data07-03.sav	474 名银行雇员数据	172
data07-04.sav	某公司不同性别经理薪金情况	181
data07-05.sav	地产评估数据	184
data07-06.sav	肺癌患者生存时间数据	186
data07-07.sav	200 例正常人血铅含量	187
data07-08.sav	150 名 3 岁女童身高数据	187
data07-09.sav	6400 人生活状况好现代化工具使用调查数据	188
data07-10.sav	不同质量等级（标准、高级）合金形成温度	188
data08-01.sav	学生身高与体重	194
data08-02.sav	120 名 12 岁男孩身高	197
data08-03.sav	29 名 13 岁男生身高、体重、肺活量数据	201
data08-04.sav	体育疗法对高血压患者疗效	203
data08-05.sav	方便面饼重量抽检数据	204
data08-06.sav	减肥训练效果数据	204
data08-07.sav	两培训中心标准化考试数据	204
data08-08.sav	银行雇员工资学历等数据	199, 200
data09-01.sav	不同饲料比较数据	210, 215
data09-02.sav	不同细菌对三叶草含氮量的影响	217
data09-03.sav	四个种系雌鼠子宫重量	226
data09-04.sav	药物对红细胞增加作用	228
data09-05.sav	不同土壤对甜菜产量影响	230, 231
data09-06.sav	镉作业工人肺活量与年龄、接触时间数据	234
data09-07.sav	教育心理学研究数据	235, 265
data09-08.sav	1481 个心梗患者的数据	240
data09-09.sav	刺激与反应时测量数据	253
data09-10.sav	四种药物对某生化指标的作用（重复测量设计）	256
data09-10a.sav	data09-10.sav 数据结构重建	258
data09-11.sav	两种记忆方法的比较	259
data09-12.sav	航空公司、零售业、旅馆业和汽车制造业评定数据	268
data09-13.sav	银行雇员基本情况和工资数据	268
data09-14.sav	三种麻醉诱导方法在不同时相测量的收缩压变化	268
data10-01.sav	安徽省国民收入与城乡居民储蓄存款余额	272
data10-02.sav	474 个银行雇员数据	273, 274, 285
data10-03.sav	前 10 名运动员长拳和长兵器两项得分	275
data10-04.sav	四川绵阳地区中山柏生长与自然环境关系资料	277, 283, 284
data10-05.sav	身高、体重、肺活量数据	285
data10-06.sav	太阳镜销售情况	285
data11-01.sav	474 名银行雇员工资数据	296, 300, 308

数 据 编 号	数 据 名 称	出 现 页 码
data11-02.sav	1969—1971 年 美国某银行 474 雇员状况数据	308, 312
data11-03.sav	汽车数据	315
data11-04.sav	乳腺癌细胞淋巴转移数据	322
data11-05.sav	1992 年美国总统大选得票结果	331
data11-06.sav	不同婚姻状态的幸福感数据	340
data11-07.sav	鱼藤酮杀虫剂浓度与杀虫量数据	344
data11-08.sav	美国 1790—1960 年人口数据	351, 352
data11-09.sav	教学实验数据	354, 355
data11-10.sav	80 名不同受教育程度员工的工资数据	360, 361
data11-11.sav	1991 年美国社会情况调查	378
data11-12.sav	美国赖特州立大学医学院对千名高中学饮酒方面的调查数据	385, 389
data11-13.sav	相信死后有来世的调查结果	391, 392, 399
data11-14.sav	某企业 1987—1998 年科研经费与经济效益数据	404
data11-15.sav	某商场 1989—1998 年商品流通过费率与商品零售额	404
data11-16.sav	电流刺激农场动物的实验数据	404
data12-01.sav	300 次掷一颗六面体实验观测结果（原始录入方式）	408
data12-01a.sav	data12-01.sav（频数录入方式）	408
data12-02.sav	100 名健康成年女子血清总蛋白含量（原始录入方式）	409
data12-02a.sav	data12-02.sav（频数录入方式）	409, 447
data12-03.sav	31 次掷一枚比赛挑边器实验观测结果（原始录入方式）	411
data12-03a.sav	data12-03.sav（频数录入方式）	411
data12-04.sav	掷硬币数据	413
data12-05.sav	质点数与时间间隔数据（原始录入方式）	415
data12-05a.sav	data12-05.sav（频数录入方式）	415
data12-06.sav	两种安眠药效果对比	419
data12-07.sav	4 种不同操作方法优等品率实验数据	422, 439
data12-08.sav	锻炼前后晨脉比较	425
data12-09.sav	顾客对 3 种款式的衬衣的喜爱程度	427
data12-10.sav	100 名健康成年女子血清总蛋白含量数据	434
data12-11.sav	村长选举中对 50 位村民中的摸底调查数据	445
data12-12.sav	设备进行寿命试验，记录 10 次无故障工作时间	447
data12-13.sav	监听装置接收信号实验数据	447
data12-14.sav	两个地点的地表土壤 pH 值	447
data12-15.sav	某种药物治疗前后血压变化	447
data12-16.sav	20 个村民对 4 位候选人满意度调查	447
data13-01.sav	汽车性能指标与销售数据	453
data13-02.sav	10 名游泳运动员的三项测试	458, 459
data13-02a.sav	data13-02.sav（作为初始聚类中心的种子数据）	459, 460
data13-02b.sav	聚类结果的类中心--新种子数据文件	459
data13-03.sav	20 种啤酒数据	467, 474
data13-04.sav	有关学生 10 个测验项目数据	475
data13-05.sav	鸢尾花分类与特征	484, 492
data13-06.xls	74 个国家人口出生率和死亡率（Excel 格式数据）	496
data13-07.xls	标枪运动员等级成绩数据（Excel 格式数据）	496
data13-07a.sav	标枪运动员等级。素质成绩数据	496
data13-07b.sav	待判等级的标枪运动员素质数据	496
data14-01.sav	标准大城市人口调查区的 5 个经济学变量数据	502, 507

数据编号	数据名称	出现页码
data14-01a.sav	data14-01.sav (带有因子分数变量)	510
data14-01b.sav	data14-01a.sav 排序后的结果	514
data14-02.sav	顾客对车型偏好的研究调查	514
data14-02a.sav	data14-02.sav (带有因子分数好聚类结果)	516
Data14-02b.sav	data14-02.sav (带有因子载荷)	516
data14-03.sav	10 个省份主要消费支出比例	521
data14-04.sav	某医院 3 年的治疗与经营数据	524
data14-05.sav	31 个省市自治区各种经济类型资产占总资产比重	524
data15-01.sav	研究运动员意志品质的调查	529, 530
data15-02.sav	顾客对饮料相似性感受的调查	534
data15-03.sav	受试者对牙膏品牌的认识数据	536
data16-01.sav	酸奶调查设计文件	550
data16-02.sav	带有两个模拟观测的酸奶调查设计数据	558
data16-03.sav	地毯清洁器的调查设计	561, 562
data16-04.sav	地毯清洁器调查数据	564
data17-01.sav	某公司 1973—1999 年的销售额	573
data17-02.sav	85 地区宽带供应商 1999 年 1 月—2003 年 12 月服务用户数量	575, 601, 604
data17-03.sav	Data17-02 补进 2004 年 1—3 月份数据	600
data17-04.sav	某公司 1986—1997 年各季度商品销售数据	604, 609, 614
data17-05.sav	某国际航线 1949 年 1 月—1960 年 12 月的旅客数据	618
data17-06.sav	1989 年 1 月—1998 年 12 月三种男女服装产品销售情况	621
data17-07.sav	某邮购公司 1989 年 1 月—1998 年 12 月服装销售及宣传等数据	622
data18-01.sav	不同饮食下 90 只老鼠的无肿瘤时间	628
data18-02.sav	58 例肾上腺样瘤患者不同治疗方式下的生存时间	633
data18-03.sav	137 位肺癌患者生存时间	638
data18-04.sav	SPSS 自带数据文件 recidivism.save 中的数据	641
Data18-05.sav	3 期和 4 期黑瘤患者的数据	644
Data18-06.sav	63 例患者的生存时间、结局及影响因素	645
data19-01.sav	我国 12 个城市 1985—1994 年每城市月平均气温	646, 653, 654, 657
data19-02.sav	数据同 data20-01.sav, 不同文件结构	646
data19-03.sav	1985—1994 年上海月平均气温	646
data19-04.sav	1988—1992 年世界各种饮料产量	647, 650
data19-05.sav	451 青少年生理形态数据	650, 659-663
data19-06.sav	不同时期不同性别毕业生初始薪酬	651
data19-07.sav	1950—1985 年我国国防、经济支出	653
data19-08.sav	1996.4.1—1996.4.19 上证所地产类股票价格	654
data19-09.sav	1996.4.1—1996.4.19 上证所北京地区股票价格	655
data19-10.sav	1996.4.1—1996.4.19 上证所几支工业和商业股票价格	656
data19-11.sav	某年部分独联体国家失业人口数据	657
data19-12.sav	银行雇员工资数据	658
data19-13.sav	150 名 3 岁女童身高	664
data19-14.sav	200 名正常人血铅含量	664
data19-15.sav	某刀具厂切削刀次品件数分类计数	665
data19-16.sav	各国医疗保健从业人数	665
data19-17.sav	汽车空调蒸发器故障分类及计数	666
Data19-18.sav	各国制造加工业雇佣人数 100 人以上工厂数量	666
Data19-19.sav	男女各年龄司机每百万公里伤亡和非伤亡事故数据	667, 668

数 据 编 号	数 据 名 称	出 现 页 码
Data19-20.sav	1988 年 6 月 1 日—30 日每日早中晚三班电解工序的电解效率(1)	668
Data19-21.sav	1988 年 6 月 1 日—30 日每日早中晚三班电解工序的电解效率(2)	668
Data19-22.sav	某搅拌站实测混凝土坍落度数据	670
Data19-23.sav	某种小螺钉检测数据	671
Data19-24.sav	某医院每月出现危急外科手术例数	673
Data19-25.sav	某轧钢厂生产的 6mm±0.4mm 厚度钢板测试记录	669
Data19-26.sav	某构件厂产品质量数据	671, 672, 673
Data19-27.sav	抽样数不等的小螺丝检测数据	671, 672, 673
Data19-28.sav	世界人口数据	673
data20-01.sav	我国 12 城市平均气温	667, 680
data20-02.sav	451 名青少年体质数据	692, 695
data20-03.sav	某年部分独联体国家失业人口数据	697
data20-04.sav	世界各国饮料产量	691

# 参 考 文 献

- [1] George A. Morgan, Orlando V. Griego. Mahwah. Easy use and interpretation of SPSS for Windows: answering research questions with statistics. NJ Lawrence Erlbaum, c1997.
- [2] Duncan Cramer. Introducing statistics for social research: step-by-step calculations and computer techniques using. SPSS. London Routledge, 1994.
- [3] SAS/BASE guide for Personal Computer. SAS Institute Inc, 1988.
- [4] SAS/STAT Guide for Personal Computer. SAS Institute Inc, 1988.
- [5] SPSS Base 7.5 for Windows user's guide. SPSS Inc, 1997.
- [6] SPSS graphics. SPSS Inc, 1985.
- [7] Marija J. Norusis. SPSS professional statistics 6.1. Chicago IL, 1994.
- [8] Naresh K Malhotra. Marketing Research. 市场调研. 北京: 清华大学出版社, 1998 年第 1 版
- [9] 卢纹岱, 金水高. SAS/PC 统计分析实用技术. 国防工业出版社, 1996.
- [10] 高惠璇, 张庆峰等编译. SAS 系统与市场调查数据分析. 北京大学概率统计系, 1997.
- [11] 吴明隆. SPSS 系统应用务实. 北京: 中国铁道出版社, 2000 年第 1 版.
- [12] 汪贤进. 常用统计方法手册. 杭州: 浙江人民出版社.
- [13] Douglas M Bates. 非线性回归分析及其应用. 北京: 中国统计出版社, 1998.
- [14] D.A.Ratkowsky. 非线性回归模型. 南京: 南京大学出版社, 1986.
- [15] 郝德元. 教育与心理统计. 北京: 教育科学出版社, 1982.
- [16] Elisa T Lee. 生存数据分析的统计方法. 北京: 中国统计出版社.
- [17] 孙尚拱. 实用多变量统计方法. 北京: 中国医科大学与中国协和医科大学联合出版社, 1990.
- [18] 吴国富. 实用数据分析方法. 北京: 中国统计出版社.
- [19] 袁淑君. 数据统计分析——SPSS/PC<sup>+</sup>原理及其应用. 北京: 北京师范大学出版社, 1995.
- [20] 周兆麟. 数理统计学. 北京: 中国统计出版社, 1987.
- [21] 张元. 田间实验与生物统计. 沈阳: 东北师大出版社, 1986.
- [22] 贾宏宇. 统计辞典. 上海: 上海人民出版社, 1986.
- [23] 郑家亨. 统计大辞典. 北京: 中国统计出版社.
- [24] David F Freedmen. 统计学. 北京: 中国统计出版社, 1997.
- [25] 胡学锋. 统计学. 广州: 中山大学出版社, 1999.
- [26] 黄德霖. 统计学. 北京: 人民日报出版社, 1988.
- [27] 杨树勤. 卫生统计学. 北京: 北京人民卫生出版社, 1993.
- [28] 胡良平. 现代统计学与 SAS 应用. 北京: 军事医学科学出版社, 1996.
- [29] 方积乾. 医学统计学与电脑实验. 上海: 上海科学技术出版社, 1997.
- [30] 史秉璋. 医用多元分析. 北京: 北京人民卫生出版社, 1988.
- [31] 金丕焕. 医用统计方法. 上海: 上海医科大学出版社, 1992.
- [32] 贾怀勤. 应用统计学. 北京: 对外贸易教育出版社.

- [33] S Weisberg. 应用线性回归. 北京: 中国统计出版社, 1998.
- [34] 吴辉. 英汉统计词汇. 北京: 中国统计出版社, 1987.
- [35] 杨树勤. 中国医学百科全书——医学统计学. 上海: 上海科学技术出版社, 1985.
- [36] 最新质量统计技术及其应用. 北京: 机械工业出版社.
- [37] 柯惠新, 丁立宏编著. 市场调查与分析, 北京: 中国统计出版社, 2000 年 3 月.
- [38] 郑日昌, 蔡永红, 周益群. 心理测量学. 北京: 人民教育出版社, 1999 年 9 月第 1 版.
- [39] 袁淑君, 孟庆茂. 数据统计分析——SPSS/PC<sup>+</sup>原理及其应用. 北京: 北京师范大学出版社, 1995 年 2 月第 1 版.
- [40] 谢小庆. 信度估计的系数[J]. 心理学报, 1998(30). 2: 193—196.
- [41] 侯杰泰. 信度与度向性: 高 Alpha 量表不一定是单向度[J]. 教育学报(香港), 1995(23), 1:142.
- [42] R.A.Johnson, D. W.Wichern 著, 陆璇译实用多元统计分析(第四版), 北京: 清华大学出版社. 2001 年 4 月.
- [43] 孙振球, 徐勇勇. 医学统计学. 北京: 人民卫生出版社, 2002 年 8 月.
- [44] Naresh K.Malhotra. 市场调查(第二版). 北京: 清华大学出版社, 1998 年 8 月第 1 版.
- [45] David Freedman. 统计学. 北京: 中国统计出版社, 1997.
- [46] S.Weisberg. 应用线性回归. 北京: 中国统计出版社, 1998.
- [47] Douglas M.Bates. 非线性回归及其应用. 北京: 中国统计出版社, 1997.
- [48] 胡学锋. 统计学. 广东: 中山大学出版社, 1999.
- [49] SPSS Advanced Models 10.0. USA: SPSS Inc, 2000.
- [50] SPSS Regression Models 10.0. USA: SPSS Inc, 2000.
- [51] 阮桂海等. SPSS for Windows 高级应用教程. 北京: 电子工业出版社, 1998.
- [52] 孙明玺. 预测和评价. 浙江: 浙江教育出版社, 1986.
- [53] 于秀林, 任雪松. 多元统计分析. 北京: 中国统计出版社, 1998.
- [54] 徐国祥. 统计预测和决策. 上海: 上海财经大学出版社, 1998.
- [55] George E.P.Box. 时间序列分析预测与控制. 北京: 中国统计出版社, 1997.
- [56] 吴喜之. 非参数统计. 北京: 中国统计出版社, 1999.
- [57] 张建华, 王健等译. 商务与经济统计(第七版). 机械工业出版社, 2000 年 4 月第 1 版.
- [58] 陈鹤琴, 罗明安译. 例解商务统计. 北京: 清华大学出版社.